numero 111
novembre 1977
anno X
volume XIX

LE SCIENZE
edizione italiana di SCIENTIFIC AMERICAN

L'epatite virale

Molto si è appreso in questi ultimi anni sulla immunologia di questa grave malattia e presto dovrebbe essere disponibile un vaccino efficace per una delle sue forme più diffuse, chiamata comunemente epatite B

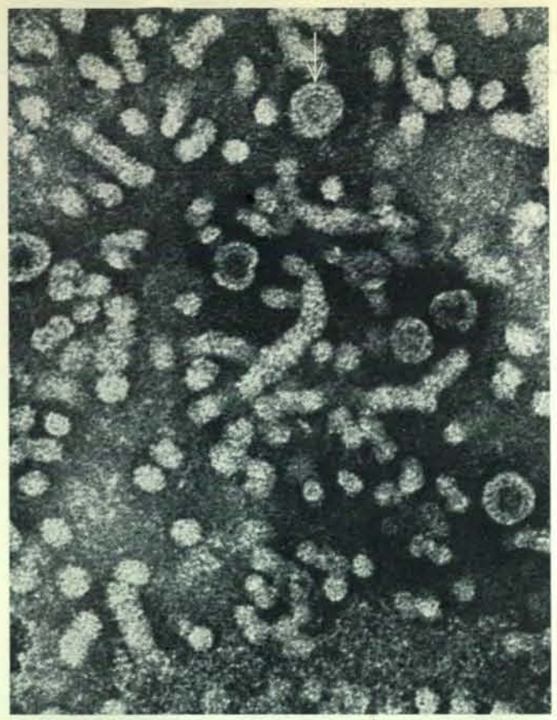
di Joseph L. Melnick, Gordon R. Dreesman e F. Blaine Hollinger

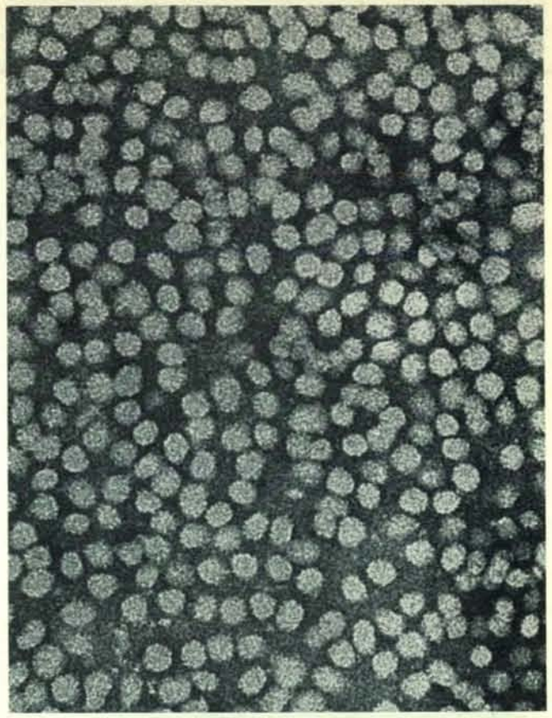
Fattore indispensabile per la produzione di un vaccino contro una malattia virale è la disponibilità di una grossa quantità di virus e in questo secolo il problema cruciale dei ricercatori che lavorano alla produzione di vaccini virali è stato quello di scoprire un valido sistema di coltivazione del virus, come su uova embrionate o su colture cellulari. Va sottolineata quindi l'importanza che siano attualmente in corso di sperimentazione vaccini che potrebbero essere in grado di tenere sotto controllo una grave malattia come l'epatite *B*, nonostante non sia mai stato possibile coltivare in laboratorio il virus che provoca tale malattia. I vaccini sono costituiti da componenti di quello che sembra essere un involucro o guscio esterno del virus. Non si conosce né la funzione precisa di tali componenti né il momento e il meccanismo con cui essi vengono a trovarsi associati nella particella virale. Tuttavia essi compaiono in gran numero nel sangue dei pazienti affetti da epatite *B* e dei portatori sani di tale malattia. Stimolando la produzione di anticorpi altamente specifici essi inducono l'immunità anche contro il virus infettivo stesso. In questo resoconto di carattere generale sull'epatite virale ci sforzeremo di ripercorrere la via che ci ha portati vicino alla scoperta di un vaccino contro l'epatite *B*.

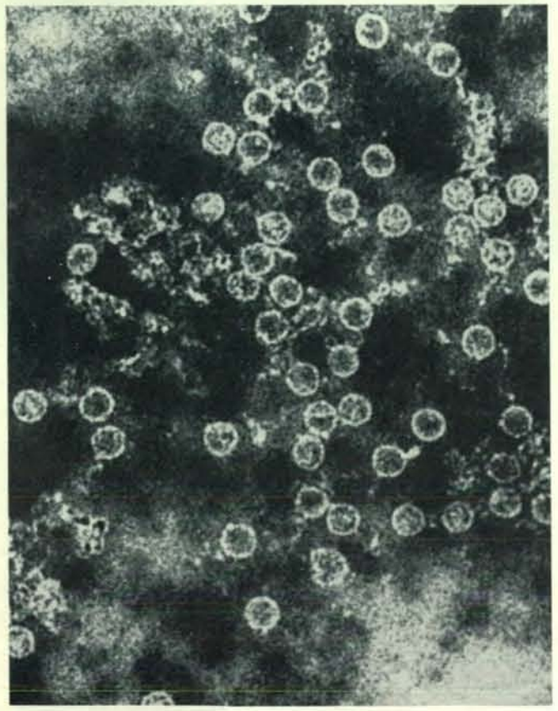
Epatite virale è un termine che viene generalmente riservato alle infezioni epatiche acute provocate da uno di almeno tre differenti virus. Il sintomo più facilmente rilevabile di queste malattie è l'ittero cioè una colorazione giallastra della cute e della sclera dovuta alla deposizione di pigmenti biliari che il fegato leso non è più in grado di rimuovere dal sangue. Una delle forme più diffuse di epatite virale è quella di tipo *A* o epatite a breve incubazione, più comunemente chiamata epatite infettiva (o ittero epidemico, o ittero catarrale acuto). Il virus che la provoca viene generalmente trasmesso con l'ingresso di alimenti contaminati dalle feci di una persona infetta nel tubo digerente di un soggetto ricettivo, ma può venir trasmesso anche per altre vie. L'altra forma più diffusa è l'epatite virale di tipo *B* o epatite a lunga incubazione, più comunemente conosciuta come epatite da siero (ittero da siero omologo, ittero da trasfusione o epatite da «siringa»). Come è implicito nelle prime denominazioni, l'epatite *B* viene trasmessa con l'inoculazione percutanea di sangue o di componenti ematici infetti; tuttavia studi recenti dimostrano che il virus si può trasmettere anche per altre vie. Oltre a questi due tipi di infezione ben definiti, è stata individuata una nuova forma di epatite associata alle trasfusioni di sangue, somigliante all'epatite *B* per quanto riguarda i sintomi clinici e il tempo di incubazione. Essa sembra provocata da uno o più agenti non correlati ai virus dell'epatite *A* o *B*, e per comodità viene chiamata «epatite *C*», anche se effettivamente sembrano esservi implicati molti agenti diversi.

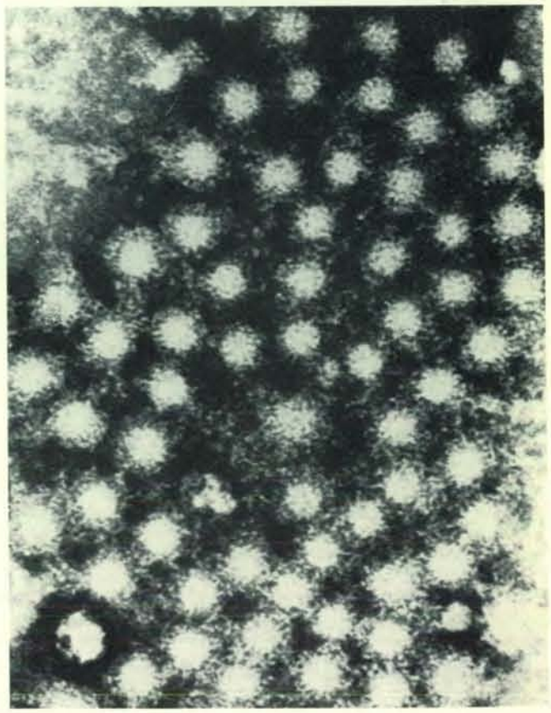
L'anno scorso sono stati registrati negli Stati Uniti circa 60 000 casi di epatite acuta di cui circa il 60 per cento furono diagnosticati come epatite *A*, il 25 per cento come epatite *B* e nel 15 per cento dei casi non fu specificato il tipo. Molti casi sono, però, così lievi da passare inosservati e i medici spesso non segnalano la malattia anche quand'essa viene diagnosticata; il numero reale dei casi è quindi stato stimato circa 10 volte superiore al numero segnalato. Sulla base di questi dati l'epatite virale è quindi da considerare una malattia importante dal punto di vista della spesa pubblica e dell'inabilità individuale. Per quanto riguarda l'epatite *B*, alcuni pazienti a guarigione clinica avvenuta, diventano portatori del virus, mentre altri sviluppano la malattia in forma cronica. Negli Stati Uniti vi sono circa un milione di portatori di epatite *B*.

Fino all'inizio degli anni sessanta i due differenti tipi di virus responsabili dell'epatite *A* e dell'epatite *B* non erano stati caratterizzati; si sapeva semplicemente che sembravano esistere due forme differenti di malattia, una con periodo di incubazione più lungo dell'altra. In seguito Saul Krugman, della New York University School of Medicine, e i suoi colleghi raccolsero e conservarono il siero di pazienti affetti da una delle due forme di malattia. Essi osservarono che i bambini esposti al siero di pazienti affetti dalla forma a breve incubazione manifestavano regolarmente un'epatite clinicamente evidente circa 30 giorni dopo, mentre bambini inoculati con siero appartenente a pazienti affetti dalla forma a lunga incubazione sviluppavano la sintomatologia dopo 60 giorni. Il siero proveniente da questi nuovi pazienti trasmetteva a sua volta la forma specifica della malattia nei bambini a cui veniva somministrato. Il superamento di una delle due forme di epatite conferiva immunità solo nei confronti di quel tipo di

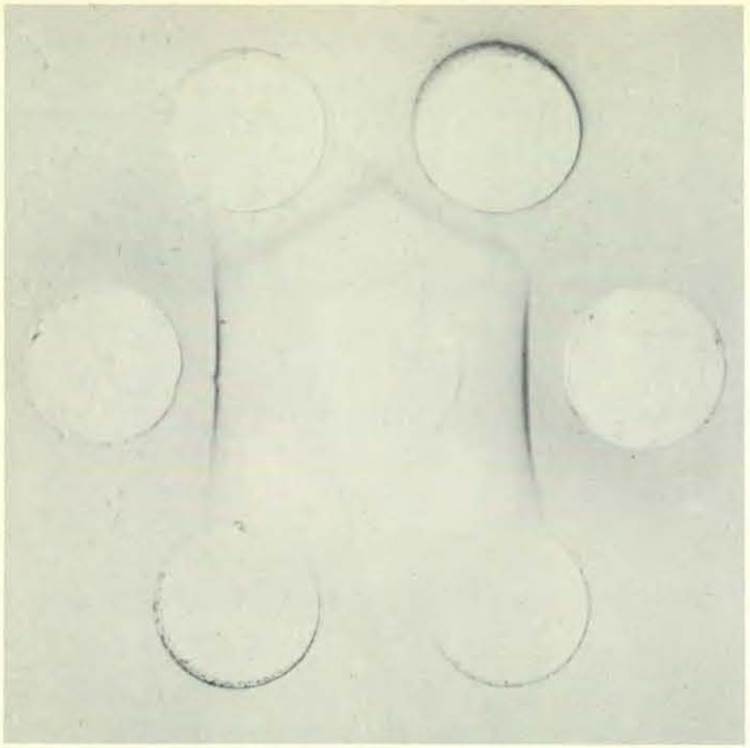








Il virus dell'epatite *B* e le sue subunità sono ingrandite 200 000 volte in tre microfotografie elettroniche ottenute da Guy A. Cabral nel laboratorio degli autori. Tre particelle principali si osservano nel sangue di pazienti affetti da epatite *B* (*in alto a sinistra*): piccole sfere di circa 22 nanometri di diametro, strutture filamentose con diametro trasversale di 22 nanometri e la forma completa a doppio involucro del virus dell'epatite *B*. Le numerose sfere da 22 nanometri possono essere purificate (*in alto a destra*) e rappresentano il materiale principale utilizzato per i vaccini. Le cellule epatiche dei pazienti affetti da epatite *B* contengono particelle grandi 26 nanometri (*in basso a sinistra*), che sono la struttura centrale non rivestita del virus a doppio involucro. Nella microfotografia elettronica di un «aggregato immune», in cui le particelle virali sono tenute assieme da anticorpi specifici, eseguita da Clifton R. Gravelle dei laboratori Phoenix del Center for Disease Control, il virus dell'epatite *A*, che non ha niente a che fare con il virus dell'epatite *B*, è ingrandito 200 000 volte (*in basso a destra*).



Il metodo della gel-diffusione, elaborato da Orjan Ouchterlony del Karolinska Institute di Stoccolma, è in grado di evidenziare l'antigene di superficie dell'epatite *B* nel siero e permette inoltre di dimostrare che esistono differenti sottotipi di tale antigene. Sette pozzetti sono stati ricavati in uno strato di agar steso su di un vetrino (ingrandito qui di circa sette volte). Il siero contenente l'anticorpo diretto contro l'antigene di superficie è stato posto nel pozzetto centrale; il plasma di sei individui in esame è stato seminato nei pozzetti periferici. Gli antigeni presenti nei sieri e l'anticorpo presente nel pozzetto centrale, diffondendosi attraverso il gel, formano una linea di precipitato nel punto di incontro. Questo esempio di gel-diffusione secondo Ouchterlony indica che quattro dei sei sieri contengono l'antigene di superficie, che in tre casi appartiene a uno stesso sottotipo e in un caso a un sottotipo diverso. Il meccanismo molecolare mediante il quale si formano le linee di precipitato è spiegato nell'illustrazione della pagina a fronte.

malattia e non dell'altro. Krugman e Joan P. Giles dimostrarono che se il siero di un paziente affetto da epatite *B* veniva inattivato con l'ebollizione e poi inoculato, i bambini che avevano ricevuto siero inattivato col calore risultavano protetti dall'infezione provocata dal virus dell'epatite *B*. Tali esperimenti dimostravano che due differenti virus erano responsabili di due differenti forme di epatite. Essi indicavano anche che sarebbe stato possibile produrre vaccini contro le malattie qualora i virus fossero stati isolati o coltivati.

La catena di eventi che condusse alla scoperta del virus dell'epatite *B* fu però complessa. Cominciò nel 1963 quando Baruch S. Blumberg dell'Institute for Cancer Research di Filadelfia intraprese l'esame di migliaia di campioni di sangue di popolazioni diverse per uno studio non sull'epatite, ma sulla variazione genetica delle proteine seriche. Egli scoprì che un campione di siero di un aborigeno australiano conteneva un antigene che reagiva con un anticorpo presente nel siero di un paziente americano affetto da emofilia. Questo paziente che aveva ricevuto numerose trasfusioni di sangue, aveva molto probabilmente prodotto un anticorpo in risposta ad una precedente esposizione all'«antigene Australia». Che cos'è questo antigene? Studi successivi hanno dimostrato che esso è molto raro nelle popolazioni dell'America settentrionale, mentre è molto più comune in alcune popolazioni africane e asiatiche. Blumberg e i suoi collaboratori scoprirono un legame tra questo antigene e la malattia che ora conosciamo come epatite da siero. Nel 1968 altri ricercatori, e in particolare Alfred M. Prince del New York Blood Center e K. Okochi del Tokio University Hospital, confermarono che l'antigene Australia si trovava soprattutto nel siero di pazienti affetti da epatite *B*.

Questa importante reazione antigene-anticorpo fu scoperta e studiata inizialmente con una tecnica di diffusione su gel mediante la quale i due sieri vengono posti in pozzetti in un gel di agar, in modo che possano diffondere l'uno verso l'altro; se un siero contiene un anticorpo specifico per un antigene presente nell'altro siero, si forma una linea visibile dove l'anticorpo e l'antigene si incontrano e si combinano a formare un precipitato. È stato notato che se i sieri di due pazienti reagiscono con l'anticorpo dell'epatite *B*, a volte si forma uno «sperone» secondario che indica la presenza di più di un singolo antigene (*si vedano le illustrazioni in queste due pagine*). Le particelle virali dell'epatite *B* di ciascuna persona contengono un antigene gruppo-specifico comune agli altri sottotipi e almeno due antigeni principali sottotipo-specifici.

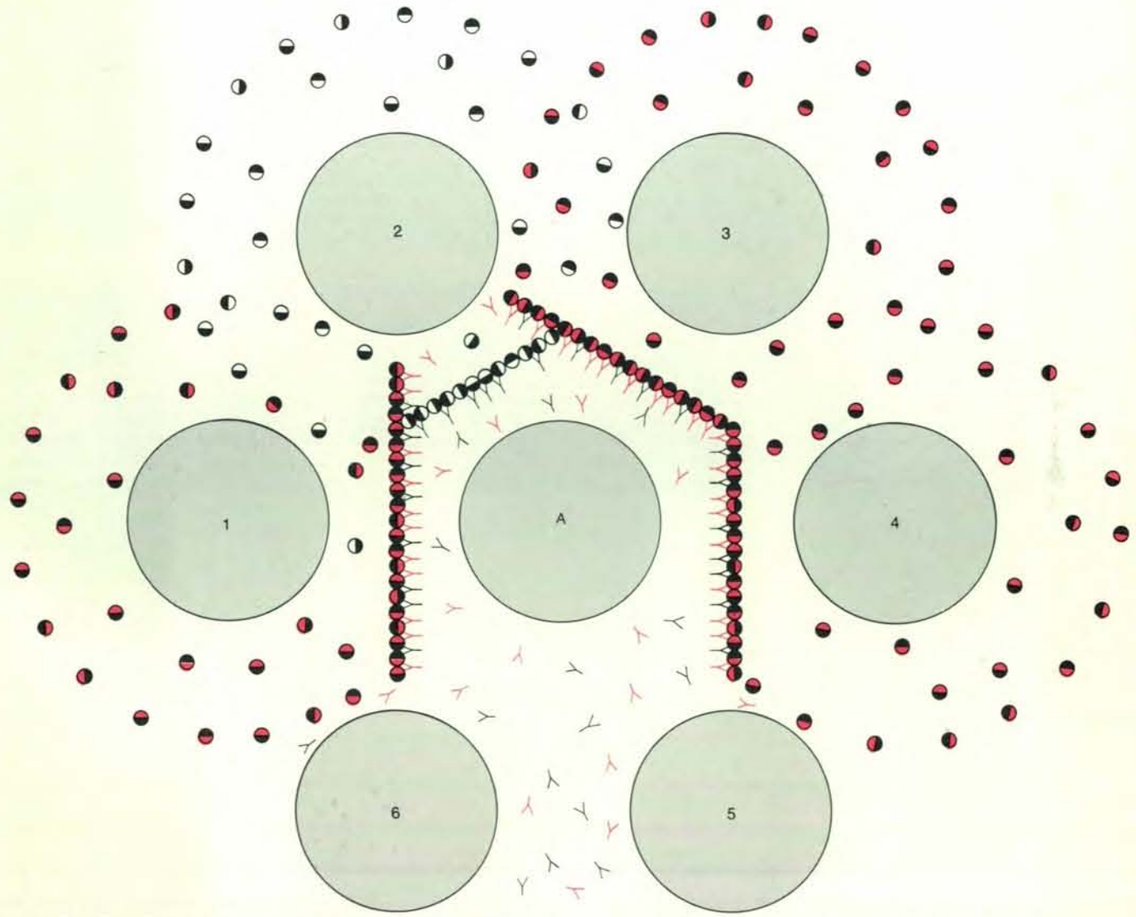
Le particelle virali dell'epatite *B* furono presto identificate: microfotografie elettroniche rivelarono che i sieri contenenti l'antigene dell'epatite *B* possedevano una varietà di particelle simili per forma e dimensioni a quelle dei virus animali noti; alcuni pazienti possedevano 1000 miliardi di particelle per millilitro di sangue. La particella virale di più frequente riscontro nella maggior parte dei campioni di sangue provenienti da pazienti affetti da epatite *B* consisteva di piccole particelle sferiche di circa 22 nanometri (milionesimi di millimetro) di diametro. C'erano anche forme lunghe e filamentose dello stesso diametro, ma con una lunghezza variante da 100 a 700 nanometri. Nel 1970 D.S. Dane della Middlesex Hospital Medical School di Londra osservò una forma ulteriore di virus: particelle simil-virali dotate di doppio guscio con un diametro esterno di circa 42 nanometri e struttura interna di 26 nanometri di diametro.

June D. Almeida dei Wellcome Research Laboratories in Inghilterra fornì il legame tra le prime osservazioni sulla reazione antigene-anticorpo e le più recenti acquisizioni basate sulle microfotografie elettroniche. Essa dimostrò che quando un siero contenente l'anticorpo contro il virus dell'epatite *B* veniva fatto reagire con un siero contenente le varie particelle antigeniche, tutte le tre particelle venivano agglutinate: le sfere di 22 nanometri, le forme filamentose e le particelle di Dane da 42 nanometri. Se i campioni di sangue venivano pretrattati con detergente per rimuovere il guscio esterno della particella di Dane, tuttavia la struttura centrale denudata non si agglutinava con le altre particelle. In altri termini, il guscio della particella di Dane era antigenicamente correlato con le sfere da 22 nanometri e con i filamenti, ma la struttura centrale della particella di Dane era antigenicamente distinta da queste forme. L'antigene associato alle sfere, ai filamenti e alla parete esterna furono designati HBsAg, o antigene di superficie dell'epatite *B*, mentre l'antigene della struttura centrale fu chiamato HBcAg. Sono state allestite metodiche molto sensibili per misurare piccole quantità di questi due antigeni. Una delle metodiche più diffuse è il test radioimmunologico in fase solida in cui il complesso antigene-anticorpo viene reso radioattivo in modo da permettere la misurazione con rivelatore di radiazioni (*si veda l'illustrazione in basso a pagina 18 e 19*); questo metodo è 4000 volte più sensibile della diffusione su gel.

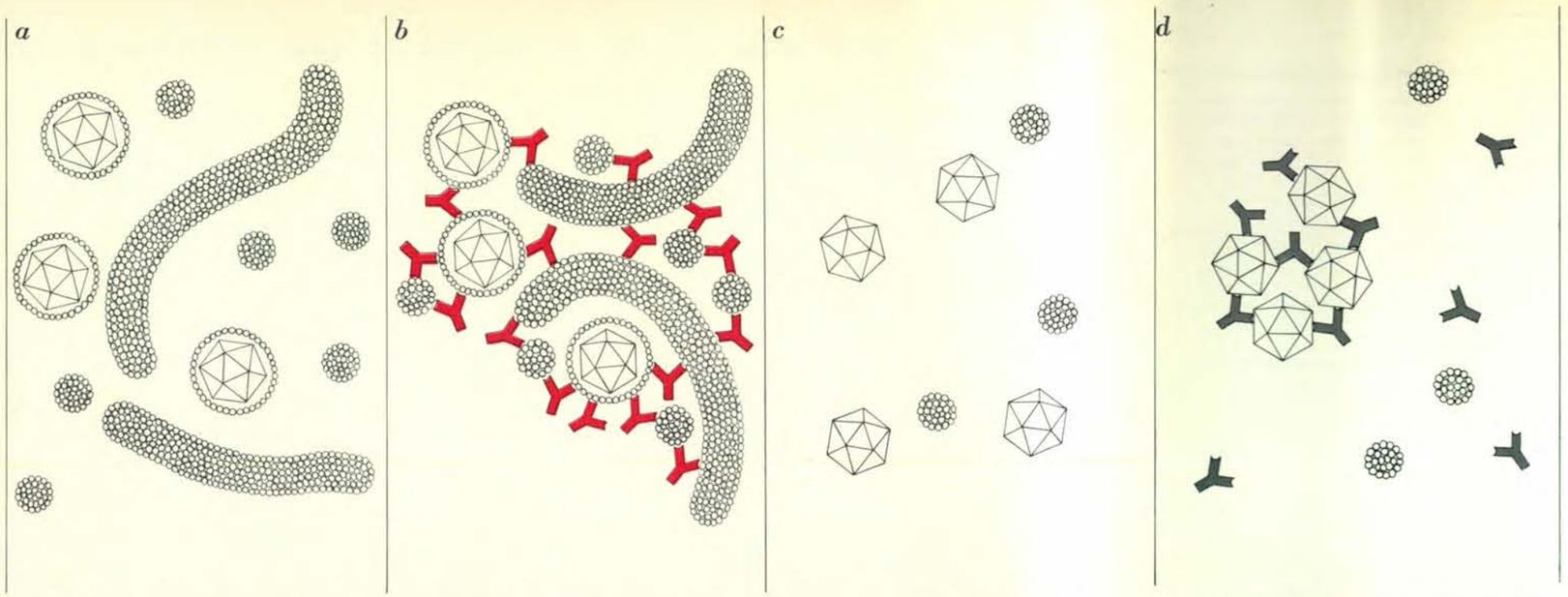
Successivamente furono purificate e caratterizzate le numerose particelle sferiche da 22 nanometri e si cercò di dimostrare se esse possedevano le caratteristiche di un virus animale. Le particelle furono purificate nel nostro laboratorio al Baylor College of Medicine e in altri laboratori con svariati procedimenti di centrifugazione e furono poi analizzate con metodi biofisici e biochimici. Noi scoprimmo che le particelle possedevano bassa densità, e ciò significava che esse contenevano sia lipidi (grassi) che proteine. Ulteriori analisi rivelarono la presenza di carboidrati ma non quella di acidi nucleici, il materiale genetico che dovrebbe essere presente in un virus. Misurazioni precise segnalarono l'esistenza di notevoli variazioni nel diametro delle particelle, da 17 a 25 nanometri. La composizione in amminoacidi delle proteine sembrava differire considerevolmente da quella dei virus animali già ben caratterizzati. Purtuttavia le particelle possedevano una struttura biochimica complessa: esse erano formate da circa otto polipeptidi diversi (catene brevi di amminoacidi) o di glicoproteine (complessi proteina-zucchero). Tutte queste osservazioni suggerivano che le particelle sferiche probabilmente non rappresentavano un virus, ma piuttosto qualche metabolita virus-specifico o materiale virale modificato dalle cellule ospiti in cui il virus si era replicato.

Poco tempo dopo, diverse particelle simil-virali furono osservate nel sangue di pazienti affetti da epatite *B* e alcuni ricercatori trovarono materiale antigenicamente correlato nelle cellule di fegato interessate dalla malattia. Il citoplasma delle cellule infettate conteneva l'antigene di superficie; l'antigene della struttura centrale fu invece ritrovato nei nuclei cellulari. Inizialmente gli antigeni furono visualizzati legando ai loro anticorpi un colorante fluorescente ben visibile alla luce ultravioletta. La loro presenza fu confermata dalla microscopia elettronica: i nuclei di epatociti infettati contenevano particelle da 26 nanometri che assomigliavano alla struttura centrale delle particelle di Dane e tali particelle non erano presenti negli epatociti di pazienti affetti da altre malattie epatiche.

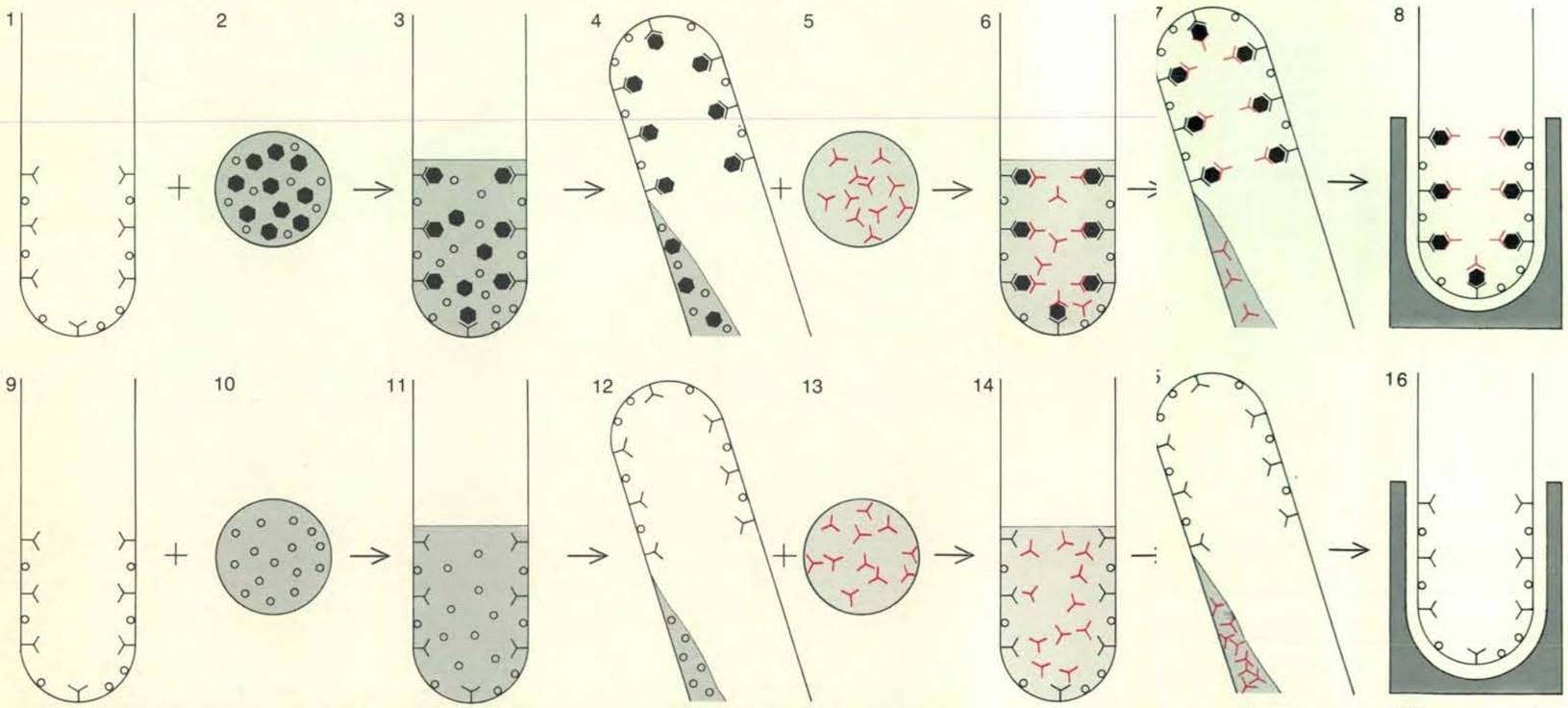
William S. Robinson e Paul Kaplan



Le linee di precipitato che si osservano nella fotografia della pagina precedente sono dovute alla reazione di legame tra antigene (*piccoli cerchi*) e anticorpo (*rappresentato dalle Y*). I sieri presenti nei pozzetti *1, 2, 3* e *4* contenevano l'antigene di superficie; i sieri dei pozzetti *5* e *6* erano negativi. Tutte le particelle dotate di attività antigenica possiedono un determinante antigenico comune «gruppo-specifico» (*in nero*), ma quelle dei pozzetti *1, 3* e *4* hanno inoltre un determinante particolare «tipo-specifico» (*in colore*), mentre quelle del pozzetto *2* possiedono un determinante «tipo-specifico» differente (*in bianco*). L'antisiero nel pozzetto *A* contiene anticorpi contro il determinante «gruppo-specifico» e contro uno solo dei due determinanti «tipo-specifici» (*quello in colore*). La linea di precipitazione tra i pozzetti *A* e *2* è più debole di quella che si è formata tra *A* e i pozzetti *1, 3* e *4* perché, in mancanza della reazione con il determinante «tipo-specifico» oltre che con quello «gruppo-specifico», si crea una quantità minore di aggregato. Tra i pozzetti *1* e *2* e tra il pozzetto *2* e *3* si possono notare due «speroni». Ciò è dovuto al fatto che l'anticorpo «tipo-specifico» (*in colore*) ha continuato a diffondere verso la periferia in assenza del corrispondente antigene nel pozzetto *2*, fino a potersi incontrare con l'antigene «tipo-specifico» dei pozzetti *1* e *3*.



Vari tipi di particelle dell'epatite *B* sono correlate o si differenziano tra loro in base all'antigenicità. Le forme che si trovano nel sangue dei pazienti sono le sfere di 22 nanometri, i filamenti con diametro trasversale di 22 nanometri e la particella virale completa (particella di Dane), che consiste di un involucro esterno e di una struttura centrale (*a*). Queste tre forme sono caratterizzate dalla presenza dell'antigene di superficie dell'epatite *B* come è indicato dal fatto che l'anticorpo (*in colore*) diretto contro questo antigene reagisce con esso provocando l'agglutinazione di ogni tipo di particella (*b*). Il trattamento con detergente elimina il rivestimento esterno che possiede l'antigene di superficie, lasciando le strutture interne «nude» (*c*). Queste a loro volta possono essere agglutinate da anticorpi diversi (*in grigio scuro*), specifici per l'antigene della struttura centrale, al fine di distinguerle dalle particelle con rivestimento (*d*).



L'antigene può essere messo in evidenza anche con la tecnica del saggio radioimmunologico in fase solida. L'anticorpo specifico per un determinato antigene (*rappresentato dalle Y in nero*) viene adsorbito su di un supporto solido, in questo caso le pareti di una provetta di plastica (*1*), assieme a proteine non specifiche (*cerchi vuoti*). Viene quindi aggiunto il materiale da studiare (*2*); se esso contiene l'antigene (*esagoni neri*), esso si combinerà con l'anticorpo (*3*). Si allontana l'eventuale eccesso di antigene (*4*) e per rivelare la presenza dell'antigene rimasto ancorato alle pareti, si aggiunge anticorpo marcato con isotopo radioattivo (*rappresentato dalle Y in colore, 5*). Esso si lega a sua volta con l'antigene, che viene a trovarsi racchiuso tra due strati di anticorpo (*6*). La provetta viene di nuovo sottoposta a lavaggio, per lasciare solo il *sandwich* anticorpo-antigene-anticorpo (*7*). La radioattività residua della provetta viene quindi misurata in un rivelatore di radiazioni (*8*) e messa a confronto con la radioattività di una provetta di controllo allo scopo di quantificare l'antigene (*9-16*).

della Stanford University School of Medicine dimostrarono che le particelle dotate di struttura centrale isolate dal sangue contenevano un DNA a doppia elica e la loro DNA-polimerasi (l'enzima necessario per la replicazione del DNA e quindi la moltiplicazione del virus). Noi abbiamo isolato le particelle dotate di struttura centrale da epatociti infettati e abbiamo dimostrato che anch'esse contenevano DNA a doppia elica e polimerasi. Il peso molecolare e la densità delle strutture centrali sono quelli tipici di certi virus animali ed esse si comportano come particelle virali quando sono sottoposte a elettroforesi. La maggior parte dei preparati di queste strutture centrali che noi abbiamo esaminato contengono tre differenti polipeptidi caratteristici del capside, o rivestimento proteico, di alcuni virus. Considerate insieme, queste diverse osservazioni erano a favore dell'ipotesi secondo cui la particella di Dane è un virus con una struttura centrale e un guscio esterno simile a quello dei virus «capsulati». L'anno scorso l'Organizzazione mondiale della sanità raccomandò infatti che la particella di Dane venisse considerata il virus dell'epatite *B*.

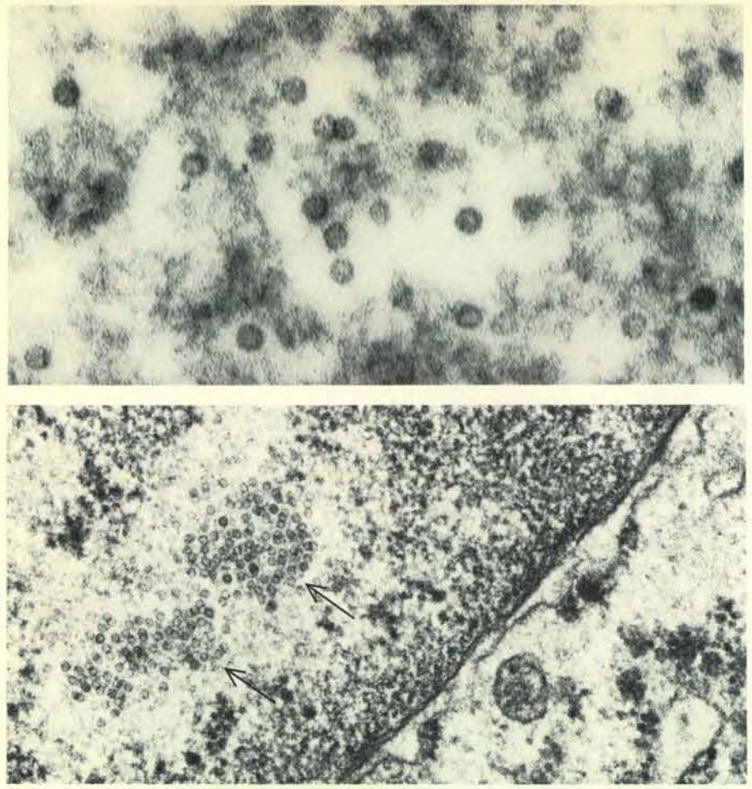
Oltre ai sistemi anticorpo-antigene di superficie e della struttura centrale, nel 1972 fu scoperto al Laboratorio batteriologico nazionale di Stoccolma un terzo sistema i cui componenti furono designati antigeni *e* dell'epatite *B* e anticorpi anti-*e*. Gli antigeni *e* sembrano essere strettamente correlati con la particella virale capsulata e il suo potere infettante. Si trova solo in individui portatori dell'antigene di superficie e la sua presenza è correlata a due altre misure del potere infettante: l'aumento della quota di particelle virali e dell'attività polimerasica dell'epatite *B*. Tra i portatori cronici dell'antigene di superficie dell'epatite *B*, quelli che possiedono l'antigene *e* sono affetti con maggior probabilità da un'epatite attiva rispetto a quelli che possiedono anticorpi anti-*e*. Allo stesso modo la trasmissione dell'epatite ai neonati è più probabile quando nel siero materno è presente l'antigene *e*, che non quando sono presenti anticorpi anti-*e*, e l'iniezione di sangue contenente antigene *e* ha maggiori probabilità di provocare l'insorgenza di epatite *B* che non l'iniezione di sangue contenente anticorpi anti-*e*. In altri termini, la presenza di antigene *e* nel sangue di portatori di antigene di superficie indica che il sangue è particolarmente infettivo; la presenza di anticorpi anti-*e* significa che il sangue è una sorgente di infezione meno probabile. Non si conosce tuttavia il locus specifico dell'antigene *e*.

Nonostante l'accumularsi dei dati riguardanti il virus dell'epatite *B* e i suoi antigeni, il virus stesso sfuggiva alle nostre ricerche e non eravamo in grado di coltivarlo in laboratorio. Fortunatamente tuttavia fu disponibile una possibilità alternativa per la produzione di vaccino. Fu dimostrato che pazienti guariti dall'epatite *B* erano protetti contro la reinfezione dalla presenza, nel loro sangue, di anticorpi diretti contro l'antigene di superficie. Studi prospettici sull'epatite provocata dalla trasfusione di sangue contenente antigene di superficie dimostrarono che nessuno dei soggetti trasfusi, il cui sangue conteneva in precedenza anticorpi contro l'antigene di superficie, si ammalò di epatite *B* clinicamente manifesta, mentre un numero significativo di trasfusi privi di anticorpi sviluppò la malattia. Al contrario la presenza o l'assenza di anticorpi contro l'antigene della struttura centrale non sembrava interferire con la protezione contro l'infezione.

Dalle osservazioni descritte sono emerse due conclusioni fondamentali. In primo luogo sembra accertato che gli anticorpi diretti contro l'antigene di superficie sono in grado di proteggere un individuo contro l'infezione prodotta dal virus dell'epatite *B*. Inoltre era evidente che le particelle antigeniche del diametro di 22 nanometri potevano essere separate dal sangue di portatori apparentemente sani in forma altamente purificata e in quantità sufficiente per la produzione del vaccino contro l'epatite *B*. Quando si scoprì che gli scimpanzè possono essere infettati dal virus dell'epatite *B*, si potè finalmente disporre di un animale da esperimento su cui saggiare tale vaccino. Come ci si aspettava, gli animali inoculati con particelle da 22 nanometri purificate e pretrattate con formaldeide produssero anticorpi contro l'antigene di superficie e non svilupparono la malattia in seguito all'inoculazione di sangue contenente virus dell'epatite *B* ad alto potere infettante. Inoltre si verificò che l'anticorpo diretto contro l'antigene gruppo-specifico conferiva un'immunità antivirale efficace, anche quando l'antigene di superficie del virus infettante era di un sottotipo differente dall'antigene di superficie del virus con cui gli animali erano stati vaccinati.

Nel dicembre scorso, Philippe Maupas dell'Università di Tours pubblicò i risultati preliminari di uno studio su volontari vaccinati con un antigene di superficie parzialmente purificato e trattato con formaldeide. Circa il 90 per cento degli individui vaccinati sviluppò una risposta anticorpale specifica verso l'antigene inoculato. Su 80 volontari vaccinati che lavoravano in un servizio di emodialisi, dove l'epatite *B* era endemica tra i pazienti e il personale, solo uno manifestò i sintomi dell'epatite *B*. Ulteriori test con vaccini costituiti da antigene di superficie sono stati condotti da Maupas e da altri ricercatori francesi.

È convinzione generale, però, che la vaccinazione con le particelle da 22 nanometri deve essere introdotta con cautela. Vi sono infatti due osservazioni che inducono alla prudenza: una riguarda la complessità dei vari componenti virali legati all'infezione da epatite *B*, complessità che potrebbe risultare maggiore di quanto ci si potrebbe aspettare semplicemente in base all'entità di informazione genetica presente nel DNA della struttura centrale del virus; l'altra è che



Nuclei di cellule epatiche di un paziente affetto da epatite *B* cronica deceduto per un'affezione renale contengono particelle nucleari del virus dell'epatite *B*. I gruppi di particelle indicate dalle frecce sono ingranditi 50 000 volte (*in alto*) mentre le particelle presenti in un altro nucleo di cellula epatica sono ingrandite 200 000 volte (*in basso*). Queste microfotografie elettroniche sono state ottenute da Ferenc Gyorkey presso il Veterans Administration Hospital di Houston.

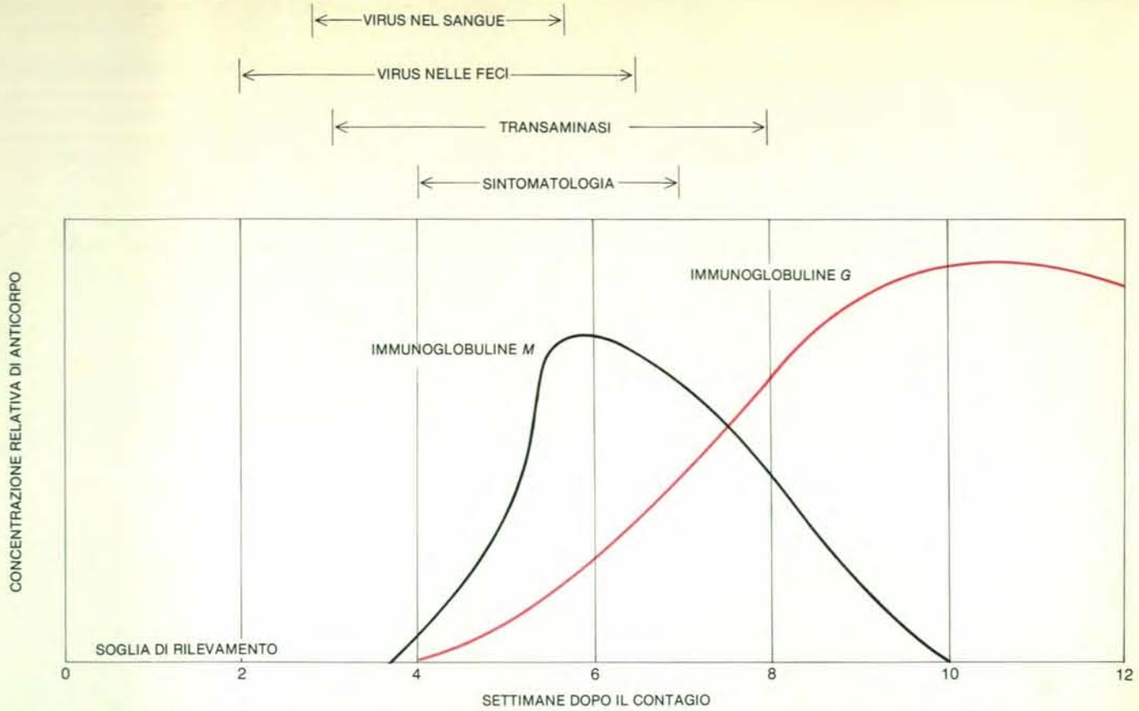
l'inoculazione nei piccoli animali da esperimento di particelle che contengono l'antigene di superficie produce una forte risposta anticorpale, mentre l'uomo (ospite naturale dell'infezione) e lo scimpanzè, perfino dopo l'infezione, producono solo bassi livelli anticorpali verso l'antigene di superficie. Queste osservazioni fanno supporre che le particelle di 22 nanometri contengano componenti cellulari antigeniche (in questo caso probabilmente proteine epatiche) derivate dall'ospite, invece di proteine «estranee» associate al virus. Conseguentemente alcuni individui potrebbero sviluppare una risposta autoimmunitaria verso le loro proprie proteine epatiche. A conferma di questa temuta conseguenza è stato segnalato che pazienti affetti da epatite cronica attiva di tipo *B* hanno sviluppato una risposta immunitaria contro lipoproteine normalmente presenti nella cellula epatica (complessi lipoprotidici).

Alla luce di questi fatti bisogna valutare la possibilità di trovare una sorgente di vaccino più sicura. Alcuni laboratori, tra cui il nostro, stanno studiando le proprietà immunitarie di componenti polipeptidici individuali delle particelle di 22 nanometri purificate. Da questi studi emerge che la maggior parte dei polipeptidi individuali induce una risposta immunitaria negli animali da esperimento e che animali inoculati con un solo polipeptide producono un anticorpo che reagisce con la particella da 22 nanometri completa. Gli animali sono stati studiati attentamente per vedere se sviluppavano una reazione autoimmunitaria verso materiale epatocellulare normale; solo i polipeptidi che non producono risposte immunitarie di questo genere potrebbero essere utilizzati per un vaccino. Preparati di polipeptidi piccoli e altamente purificati avrebbero inoltre il vantaggio di essere privi di geni virali o cellulari che potrebbero rappresentare un rischio di infezione. Qualora scoprissimo che un vaccino costituito da un singolo polipeptide ha vantaggi superiori rispetto a quello costituito dall'intera particella di 22 nanometri, dovremmo creare sistemi di produzione economici ed efficaci per l'estrazione di quantità sufficienti di polipeptide dalle particelle.

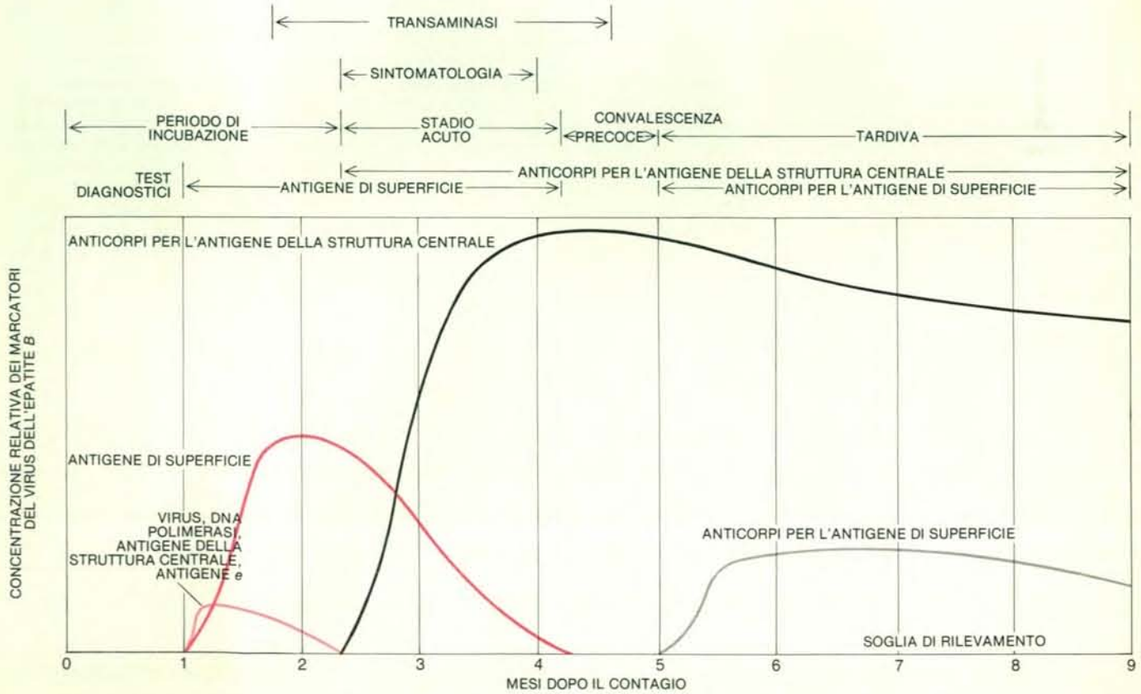
Il virus associato all'epatite *A* non è per nulla correlato a quello dell'epatite *B* e tuttora se ne sa ben poco. Esso è stato isolato per la prima volta nel 1973 da S.M. Feinstone e dai suoi colleghi ai National Institutes of Health, dalle feci prelevate durante la fase acuta della malattia da volontari sperimentalmente infettati con siero di pazienti affetti da epatite *A*. Le particelle virali hanno un diametro di 27 nanometri e, a differenza del virus dell'epatite *B*, non possiedono una capsula esterna. Il virus purificato possiede densità variabile, probabilmente perché le differenti particelle virali contengono diverse quantità di acido nucleico; alcune particelle virali sembrano essere così deficitarie da non contenere acido nucleico. Non è ancora stato possibile stabilire quale sia il materiale genetico di questi virus. Studi preliminari suggeriscono che invece del DNA, come nel virus dell'epatite *B*, l'RNA sia l'acido nucleico del virus dell'epatite *A*, come nel poliovirus e in altri virus a esso correlati. Altre caratteristiche, tuttavia, come la stabilità al calore del virus e le sue variazioni di densità, fanno supporre che il virus sia strettamente correlato a un particolare gruppo di virus di piccole dimensioni contenenti DNA. Studi conclusivi richiederanno una quantità di materiale virale dell'epatite *A* sufficiente per analisi biochimiche, il che equivale a dire che sarà necessario attendere lo sviluppo di un sistema che permetta la crescita del virus in laboratorio.

L'epidemiologia dell'epatite *A* e *B* e le variazioni seriche osservate in queste due malattie sono completamente differenti, come si verifica per infezioni causate da virus diversi. Tuttavia è molto difficile distinguere tra i due tipi di epatite sulla base della sintomatologia clinica negli stati precoci della malattia. In entrambi i casi il paziente-tipo presenta malessere, stanchezza, dolori muscolari, perdita dell'appetito, nausea, dolori addominali e febbre di basso e medio grado. L'urina diventa brunastra e la cute giallastra; a volte tutti gli altri sintomi regrediscono dopo la comparsa dell'ittero. (In alcuni casi di epatite *B* è presente uno stato infiammatorio delle articolazioni, dei vasi sanguigni e dei glomeruli renali, a volte associato a eruzioni cutanee in assenza di ittero; ciononostante, la diagnosi di epatite *B* può essere confermata da analisi biochimiche del sangue volte a dimostrare uno stato di sofferenza epatica.)

La guarigione completa si verifica entro due-quattro settimane per l'epatite *A* ed entro quattro-otto settimane per l'epatite *B*. Tuttavia circa il 10-15 per cento dei pazienti affetti da epatite *B* sviluppa una forma grave della malattia. In meno del 2 per cento di tutti i casi la malattia progredisce verso una grave e rapida diminuzione della funzionalità epatica, e il tasso di mortalità si aggira per questi casi tra il 70 e il 100 per cento. Il tasso totale di mortalità riferito a tutti i casi esaminati è dello 0,5 per cento per l'epatite *A*, del 3-4 per cento per l'epatite *B* contratta in seguito a trasfusione, dello 0,8 per cento, per l'epatite *B* associata alla tossicomania. Nel valutare queste cifre va tenuto presente che numerosi



L'epatite *A* ha un periodo di incubazione di quattro settimane, durante il quale il virus è presente nel sangue e gli enzimi epatici (transaminasi) dimostrano un livello serico aumentato. L'andamento delle curve indica la concentrazione di due tipi di anticorpi antivirus: le immunoglobuline *M*, presenti solamente nella fase acuta della malattia, e le immunoglobuline *G*, che persistono dopo la convalescenza per decenni.



L'epatite *B* ha un periodo di incubazione più lungo di quello dell'epatite *A*. Gli antigeni dell'epatite *B* e i rispettivi anticorpi compaiono, raggiungono il picco di concentrazione e iniziano a decadere in momenti diversi, condizionando l'efficacia dei test diagnostici serologici.

casi di infezione di lieve entità passano inosservati.

La terapia dell'epatite virale è prevalentemente di supporto, il che significa che non si può fare molto per controllare il decorso di questa malattia. I pazienti devono stare a riposo e astenersi dall'alcool e da altre sostanze epatotossiche. Sono stati impiegati i corticosteroidi, ma questi farmaci non si sono rivelati efficaci nell'abbreviare il decorso dei normali casi non complicati e inoltre espongono al rischio di una ricaduta se vengono sospesi troppo rapidamente. Nei casi gravi sono stati somministrati gli steroidi in associazione con concentrati di immunoglobuline anti-epatite *B*, ma tuttavia manca ancora una sicura dimostrazione della loro efficacia.

La terapia con interferon, una sostanza antivirale prodotta dalle cellule in seguito a un'infezione virale, è attualmente oggetto di studio. Sperimentazioni cliniche preliminari sono state effettuate alla School of Medicine della Stanford University con scarsi risultati in un piccolo gruppo di pazienti affetti da epatite cronica di tipo *B*. Fu osservata una temporanea scomparsa degli antigeni virali presenti nel sangue dei malati cronici senza una modificazione del quadro patologico epatico. Tali risultati suggeriscono che l'interferon può essere efficace nel limitare il potere infettante di individui portatori del virus, o anche nell'eradicare l'infezione cronica che affligge alcuni portatori della malattia, ma sono tuttavia necessari studi a «doppio cieco» su più vasti campioni sperimentali.

Gli eventi biologici e immunologici associati all'epatite *A* sono differenti da quelli legati all'epatite *B*. Nell'epatite *A* il virus prolifera probabilmente a livello delle pareti del canale intestinale, per poi raggiungere il fegato attraverso i vasi sanguigni e linfatici. L'eliminazione del virus o dei prodotti virali con le feci si verifica almeno due o tre settimane prima della comparsa dell'ittero e la malattia è più contagiosa alla fine del periodo di incubazione (prima che si manifesti la sintomatologia) quando il virus è più concentrato nelle feci, mentre esse non risultano essere contagiose trascorse tre o più settimane dalla comparsa della sintomatologia. Negli stadi precoci dell'infezione sono presenti bassi livelli serici e urinari del virus, ma si tratta tuttavia di rare sorgenti di infezione, né si pensa che la malattia si possa diffondere tramite le microgocce di vapore presenti nell'aria espirata dal naso e dalla bocca. Molti studi hanno dimostrato una relazione diretta tra la diffusione dell'epatite *A* (valutata in base alla presenza di anticorpi antivirali) e l'aumento dell'età e le scadenti condizioni igieniche ambientali e personali, correlazione comune a molte infezioni virali che interessano il tratto intestinale.

Gli eventi clinici e immunologici che seguono l'infezione da epatite *B* riflettono un'interazione complessa tra ospite, virus e antigeni virali specifici. Nei casi tipici di malattia acuta, uno o due mesi dopo l'infezione compare nel sangue lo antigene di superficie, che può persistere per pochi giorni o per molti mesi. Le particelle virali, l'antigene *e* e l'attività polimerasica compaiono nel sangue insieme all'antigene di superficie. Con il progredire dell'infezione, le cellule epatiche vanno incontro a lisi, e liberano nel sangue le transaminasi epatiche: per questo motivo la comparsa dei sintomi dell'epatite è associata a un aumento dei normali livelli di transaminasemia. Sebbene la presenza dell'antigene di superficie sia la migliore dimostrazione per una precisa diagnosi di epatite *B*, tale antigene può scomparire dal sangue in una fase insolitamente precoce; in questi casi la diagnosi può essere stabilita in base alla presenza di anticorpi diretti contro l'antigene della struttura centrale che compare insieme ai segni clinici e di laboratorio della malattia.

L'antigene di superficie persiste nel sangue dei pazienti che sviluppino un'epatite cronica o che diventano portatori asintomatici; in questi casi non si riesce a evidenziare la presenza di anticorpi diretti contro l'antigene di superficie, mentre sono dimostrabili alte concentrazioni di anticorpi specifici per l'antigene della struttura centrale. Nel sangue dei pazienti affetti da epatite cronica di tipo *B* sono spesso presenti antigene *e*, DNA-polimerasi e particelle virali, mentre nel sangue dei portatori apparentemente sani si ritrova generalmente solo l'anticorpo diretto contro l'antigene *e*. Se si riespone al virus un soggetto immune nei confronti dell'epatite *B*, si verifica una rapida risposta immunitaria di «richiamo», sostenuta da anticorpi diretti contro l'antigene di superficie che compaiono entro le due settimane successive; non è possibile dimostrare la presenza dell'antigene di superficie e il soggetto reinfettato non sviluppa la malattia.

La maggior frequenza di epatite *B* nella popolazione generalmente considerata si riscontra nel gruppo di età tra i 15 e i 29 anni, presumibilmente a causa del fenomeno della tossicomania. La frequenza è tuttavia maggiore in gruppi selezionati di individui, come i pazienti sottoposti a dialisi e il personale infermieristico addetto alla loro cura. Fino a poco tempo fa l'incidenza di epatite *B* nei soggetti trasfusi con sangue o plasma era molto alta, ma tale frequenza si è ridotta quasi del 75 per cento, poiché si è reso obbligatorio individuare i donatori infetti con test specifici quale quello radioimmunologico.

La diffusione naturale dell'epatite *B* per contatto (piuttosto che soltanto tramite puntura cutanea) sembra più comune di quanto generalmente si ritenga soprattutto nel caso di rapporti sessuali con portatori del virus (sia che i portatori abbiano una sintomatologia cronica manifesta o non) e tra coloro che appartengono a comunità affollate. Una prova del fatto che l'epatite *B* può essere trasmessa per vie che non implicano la penetrazione attraverso la cute è offerta dalla costante presenza dell'antigene di superficie in varie secrezioni corporee. Nell'ambito della famiglia, per esempio, un portatore del virus può trasmettere la malattia a chi fa uso del suo asciugamano, rasoio o spazzolino da denti. (In ciascuno di questi casi il virus potrebbe certamente penetrare attraverso la cute o le mucose.) In genere l'epatite *B* viene facilmente trasmessa da pazienti in fase acuta o da malati cronici piuttosto che da portatori asintomatici; perché un portatore sano possa trasmettere la malattia sono necessarie esposizioni ripetute.

Non tutte le persone esposte vengono infettate e tra quelle infettate solo poche sviluppano la malattia in forma clinicamente evidente. Tale percentuale dipende soprattutto dalla dose di virus penetrata e dalla via di infezione: piccole quantità di virus, o un'infezione attraverso una via diversa da quella cutanea, comportano un periodo di incubazione più lungo e una minore probabilità di sviluppare una sintomatologia manifesta, con la comparsa di ittero forse nell'un per cento dei casi solamente. Dosi maggiori e infezioni attraverso la cute possono diminuire il periodo di incubazione e aumentare la frequenza di ittero fino a un caso su cinque. Le persone infettate che non manifestano la malattia e quelle che diventano portatori (cioè coloro che mantengono livelli ematici costanti di antigene di superficie per più di quattro mesi) costituiscono importanti sorgenti di infezione.

In molti studi è stata descritta una notevole incidenza di positività per l'antigene di superficie in soggetti affetti da carcinoma epatico primitivo: l'80 per cento a Taiwan (rispetto al 15 per cento nel campione controllo) e il 35 per cento di tali pazienti in Spagna (rispetto all'un per cento nei controlli). Percentuali simili riguardanti l'antigene di superficie sono state rilevate negli Stati Uniti in pazienti con carcinoma epatico insorto in un quadro di cirrosi non alcolica preesistente. Sono state fornite diverse spiegazioni attendibili per questa associazione tra epatite *B* e carcinoma epatico. Per esempio, i pazienti con carcinoma epatico o cirrosi potrebbero essere particolarmente predisposti a contrarre l'epatite *B*, o a diventare portatori cronici, oppure l'infezione virale potrebbe condurre alla cirrosi, stato che precede l'insorgenza del tumore. Alternativamente il virus dell'epatite *B* potrebbe trasformare in senso neoplastico la cellula epatica infettata, o in altri termini, il virus sarebbe la causa diretta del tumore (condizionato forse da altri fattori, tra cui quelli genetici e ambientali). Se una delle ultime due interpretazioni è corretta, il controllo dell'epatite *B* con la vaccinazione dovrebbe ridurre anche la frequenza dei carcinomi epatici.

Nonostante l'accurata selezione del sangue dei donatori allo scopo di evitare la presenza di virus *A* e *B*, alcuni pazienti sviluppano tuttora, in seguito a trasfusione, un'epatite virale sieronegativa, senza cioè che sia dimostrabile la presenza di questi virus nel sangue di tali pazienti. Dato che attualmente non possediamo metodi capaci di identificare l'agente o gli agenti eziologici dell'epatite non-*A* non-*B* (che è responsabile ora di oltre l'80 per cento delle epatiti da siero registrate negli Stati Uniti), non si conosce molto delle sue modalità di trasmissione né delle sue caratteristiche immunologiche. È tuttavia noto che il quadro clinico e l'andamento epidemiologico sono simili a quelli dell'epatite *B*; la malattia sembra essere frequente e può comportare lo stato di portatore sano; infine questa nuova forma di epatite sembra colpire più frequentemente coloro che ricevono una trasfusione di sangue da donatori che hanno un elevato livello di transaminasi nel siero.

Finché non saranno prodotti vaccini efficaci contro l'epatite *B*, il controllo di tale infezione dovrà basarsi sui metodi tradizionali, cioè una buona igiene personale, misure generali di politica sanitaria e profilassi con immunoglobuline. Nel caso dell'epatite *A*, il mezzo preventivo più efficace consiste nel lavarsi le mani prima di mangiare o di manipolare i cibi. Per quanto riguarda l'epatite *B*, è invece fondamentale evitare il contatto con qualsiasi oggetto che possa essere contaminato da sangue, sia esso un asciugamano, un rasoio, o un ago ipodermico. La sostituzione delle banche del sangue commerciali e dei donatori a pagamento con un sistema interamente basato sull'utilizzazione di donatori volontari dovrebbe ridurre l'incidenza dell'epatite *B* e *C*. L'immunoglobulina (chiamata comunemente gammaglobulina) dovrebbe essere somministrata alle persone esposte al rischio di contrarre epatite *A* in famiglia o all'interno di un'istituzione, ma non come *routine* nelle scuole o nei luoghi di lavoro; si sa inoltre che sarebbe inutile somministrare immunoglobuline più di 6 settimane dopo l'esposizione al contagio o dopo l'insorgenza dei sintomi. Nel caso dell'epatite *B*, l'immunoglobulina può essere ottimamente impiegata quando si verificano inoculi accidentali di piccole quantità di virus. Immunoglobuline ad alta concentrazione recentemente introdotte hanno dato risultati preliminari promettenti in campioni ad alto rischio.

Per quanto riguarda la realizzazione di un vaccino efficace, le nostre conoscenze sugli antigeni dell'epatite *B*, e sul modo di prepararli, lasciano supporre che i test di vaccinazione in corso avranno esito favorevole e che entro due anni sarà possibile disporre di un vaccino per immunizzare il personale sanitario frequentemente esposto al rischio di contrarre epatite *B*. Le conoscenze sull'epatite virale che si sono accumulate negli ultimi dieci anni sono state realmente notevoli e gli accorgimenti tecnici introdotti nelle ricerche sull'epatite rappresentano nuovi strumenti per lo studio di un ampio spettro di malattie umane, di probabile origine virale, per le quali la tecnologia tradizionale non ha fornito ancora un contributo decisivo.

SONO PRONTI I RACCOGLITORI PER IL 1977 DI LE SCIENZE

Questi raccoglitori corrispondono ai volumi XVIII e XIX della rivista, e rispettivamente ai fascicoli da gennaio (n. 101) a giugno (n. 106) e a quelli da luglio (n. 107) a dicembre (n. 112), più l'indice semestrale (fornito insieme al fascicolo successivo all'ultimo numero di ciascun volume).

Sono ancora disponibili i raccoglitori dal Vol. IX al XVII, e dei raccoglitori non numerati appositamente approntati per sostituire i primi otto esauriti.

Ogni raccoglitore L. 2.250 Nuovo prezzo dal 1.1.'77

I raccoglitori si possono richiedere direttamente all'editore usando l'apposita cartolina allegata alla rivista e unendo il relativo importo; gli ordini infatti vengono evasi solo a pagamento avvenuto.

I raccoglitori si trovano anche presso i seguenti punti di vendita, sempre a L. 2.250 cadauno.

BOLOGNA: Libreria Parolini Via U. Bassi 14
FIRENZE: Libreria Marzocco Via De' Martelli 22/R
MILANO: Le Scienze S.p.A. Via Victor Hugo 2
NAPOLI: Libreria Guida A. Via Port'Alba 20/21
PADOVA: Libreria Cortina Via F. Marzolo 4
PALERMO: Libreria Dante Quattro Canti di Città
ROMA: Claudio Aranci Viale Europa 319 (EUR)
TORINO: Libreria Zanaboni C.so Vittorio Emanuele 41

Le funzioni del maggior complesso di istocompatibilità

Tra le principali funzioni attribuite a questa ristretta regione cromosomica vi è il controllo della produzione di antigeni normali e tumorali i quali hanno strette analogie con gli anticorpi umorali

di Enrico Clerici

Gli esseri umani tendono spontaneamente ad associarsi secondo principi di affinità etniche, economiche, religiose e culturali i quali, una volta accettati e codificati, vengono tenacemente difesi da ogni tentativo di modificazione dall'interno o di ingerenza dall'esterno.

Un individuo che agisca in modo tale da disturbare le regole di comportamento del gruppo al quale appartiene viene emarginato o addirittura eliminato fisicamente a secondo del grado di evoluzione o meglio, involuzione, raggiunto dalla sua società. La stessa sorte sarà riservata agli estranei che tentino di penetrare in un gruppo differente dal proprio e di violarne lo statuto.

Il razzismo, che altro non è se non quanto ora enunciato, può essere considerato come la proiezione, a livello culturale, di quanto si verifica a livello biologico in ogni singolo uomo. Infatti, l'integrità dell'organismo viene gelosamente salvaguardata dal sistema immunitario che elimina qualsiasi cellula nota sulla cui superficie compaiano, per mutazione genetica, strutture glicoproteiche, note come antigeni di istocompatibilità, differenti da quelle delle altre cellule che sono tutte uguali fra di loro in ogni individuo singolarmente considerato. Occorre precisare che cellule con identici antigeni di istocompatibilità possono svolgere funzioni differenti, così come una popolazione omogenea per la lingua parlata o per il credo religioso è pur sempre composta da individui dediti a varie attività.

L'aggressione all'organismo umano può essere, oltre che endogena (cellula mutata, di tipo tumorale), anche esogena, causata da batteri, virus o altri agenti patogeni. Anche in questo caso, la protezione dell'integrità individuale viene garantita dal sistema immunitario. È intuitivo che gli antigeni batterici o virali sono strutturalmente differenti da quelli di istocompatibilità. In certe condizioni non si tratta neppure di molecole glicoproteiche ma, per esempio, glicolipidiche come nel caso dei cosiddetti antigeni tossici o endotossine delle salmonelle che causano il tifo addominale e i paratifi. Questo fatto fa ritenere che il sistema immunitario, o meglio i linfociti che lo compongono, deve essere in grado di riconoscere sia gli antigeni di istocompatibilità delle proprie cellule, sia quelli dei batteri, dei virus o di altri agenti nocivi presenti nell'ambiente esterno e, per estensione, di distinguere tutte le strutture antigeniche proprie (*self*) da quelle non proprie (*not self*).

Ogni linfocita porta sulla sua superficie recettori capaci di riconoscere un determinato antigene. Si tratta di strutture molecolari solo parzialmente note che sono complementari a quelle dell'antigene specifico, con il quale si combinano come gli elementi di un gioco a incastri. L'unione dell'antigene con la cellula che è capace di riconoscerlo, produce un segnale di «via» grazie al quale il linfocita si divide generando una grande famiglia o clone di elementi tutti morfologicamente, ma non funzionalmente, uguali alla cellula madre. Infatti, i linfociti sensibilizzati o stimolati non si limitano più al riconoscimento passivo dell'antigene ma, poiché si tratta di un intruso, lo aggrediscono e lo eliminano per mantenere intatta l'individualità dell'organismo. Questa funzione attiva si manifesta in modo differente, a seconda che i linfociti, i quali provengono originariamente da una cellula progenitrice o staminale localizzata nei tessuti emopoietici, si siano differenziati nel timo o nella borsa di Fabrizio, o nei suoi equivalenti, a generare i cosiddetti linfociti *T* e *B*.

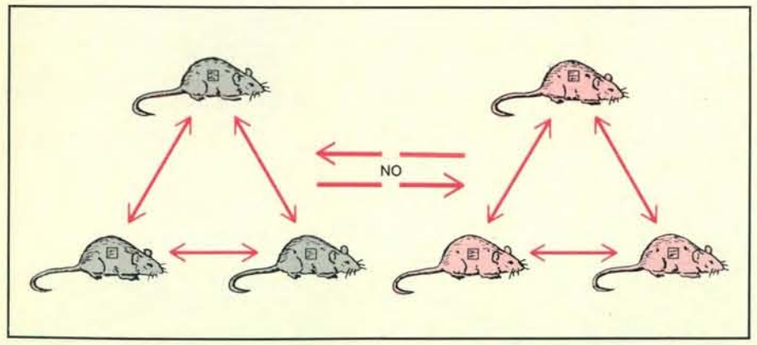
L'aggressione all'antigene viene esercitata sia dai linfociti *T* sensibilizzati, che fungono da anticorpi cellulari, sia dalle immunoglobuline sieriche. Queste immunoglobuline, più comunemente note come anticorpi umorali, sono molecole glicoproteiche che vengono sintetizzate dai linfociti *B* sensibilizzati, dopo la loro trasformazione in plasmacellule, e quindi immesse nel sangue e negli altri liquidi organici.

È opportuno, dopo queste premesse, richiamare l'attenzione sul fatto, unico in biologia, che i geni che codificano gli antigeni di istocompatibilità, nonché i recettori dei linfociti, a loro complementari, e gli anticorpi cellulari e umorali destinati ad aggredire ed eliminare gli antigeni di istocompatibilità ogni volta che per mutazione si differenzino dall'archetipo sono tutti localizzati in una ristretta regione cromosomica detta maggior complesso di istocompatibilità o *MHC*, dalle iniziali delle parole inglesi: *Major Histocompatibility Complex*. Alla conoscenza della struttura e funzioni esercitate dal maggior complesso di istocompatibilità dell'uomo si è giunti grazie agli studi, iniziati alcuni decenni orsono, sul meccanismo del rigetto dei trapianti nel topo, che è la specie di scelta per questo tipo di indagini. Come è noto, questi animali, che hanno un elevato indice di natalità, raggiungono la maturità sessuale in circa due mesi dalla nascita e la gravidanza dura solo 21 giorni. Se si ha cura di accoppiare tra loro fratelli e sorelle per almeno venti generazioni successive, si otterrà, in un tempo relativamente breve un ceppo isogenico (o singenico o *inbred*) di topi tutti geneticamente uguali fra di loro, come se si trattasse di una popolazione umana composta esclusivamente da gemelli monocoriali, ossia nati dalla fecondazione di un unico uovo da parte di un singolo spermatozoo. Quindi è come se lo sperimentatore disponesse di tante copie identiche di uno stesso animale sulle quali condurre le proprie ricerche, eliminando le interferenze dovute alla variabilità insita nei singoli componenti di una popolazione murina eterogenea. Attualmente disponiamo di centinaia di differenti ceppi isogenici i quali vengono prevalentemente utilizzati per studi di carattere immunogenetico.

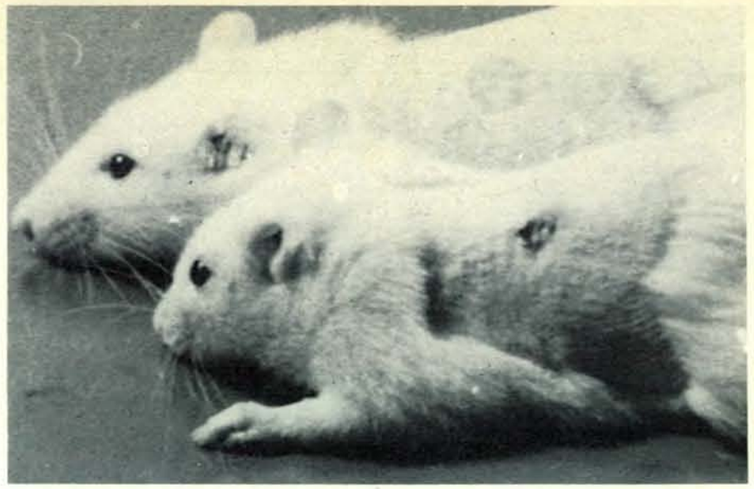
Senza entrare in dettagli, basterà dire che i trapianti di tessuto attecchiscono solo se attuati fra topi appartenenti allo stesso ceppo isogenico esattamente come avviene nel caso degli innesti eseguiti fra gemelli monocoriali umani. Invece, se un tessuto od organo di un topo di un ceppo isogenico che definiremo *A* viene innestato su un topo di un ceppo differente, che definiremo *B*, esso viene irrimediabilmente rigettato nel giro di una decina di giorni. Infatti, i linfociti del topo *B* riconoscono come estranei, non propri, gli antigeni di istocompatibilità delle cellule del topo *A* e immediatamente le eliminano. Al contrario, quando i tessuti vengono scambiati fra topi isogenici, tutti appartenenti al ceppo *A* (oppure *B*) e quindi uguali fra di loro, i linfociti dell'accettore riconoscono come propri gli antigeni di istocompatibilità del donatore e non attuano contro di essi alcuna reazione immunitaria. Questi trapianti che attecchiscono sono detti isotrapianti, mentre quelli che vengono respinti perché scambiati fra topi appartenenti a ceppi differenti prendono il nome di allotrapianti. Nel lessico immunologico si dicono poi autotrapianti gli innesti di tessuto da una parte all'altra dello stesso organismo ed eterotrapianti quelli attuati fra due specie differenti, come, per esempio, ratto su topo o viceversa.

È intuitivo che gli autotrapianti attecchiscono sempre, poiché tutte le cellule di un medesimo organismo hanno gli stessi antigeni di istocompatibilità, mentre gli eterotrapianti sono inevitabilmente destinati all'eliminazione, tanto più rapida quanto più le specie coinvolte sono tassonomicamente distanti. Questa terminologia si applica a ogni specie di vertebrato (negli invertebrati non esiste immunità specifica) incluso l'uomo, nel quale però gli isotrapianti sono limitati al solo e raro caso di quelli praticati fra gemelli monocoriali.

Gli antigeni di istocompatibilità del topo vengono prodotti da geni localizzati nel maggior complesso di istocompatibilità *H-2*, dove la lettera *H* è l'ini-



I trapianti di tessuto attecchiscono solo se lo scambio avviene tra topi dello stesso ceppo isogenico i quali, come i gemelli monocoriali umani, hanno antigeni di istocompatibilità identici.



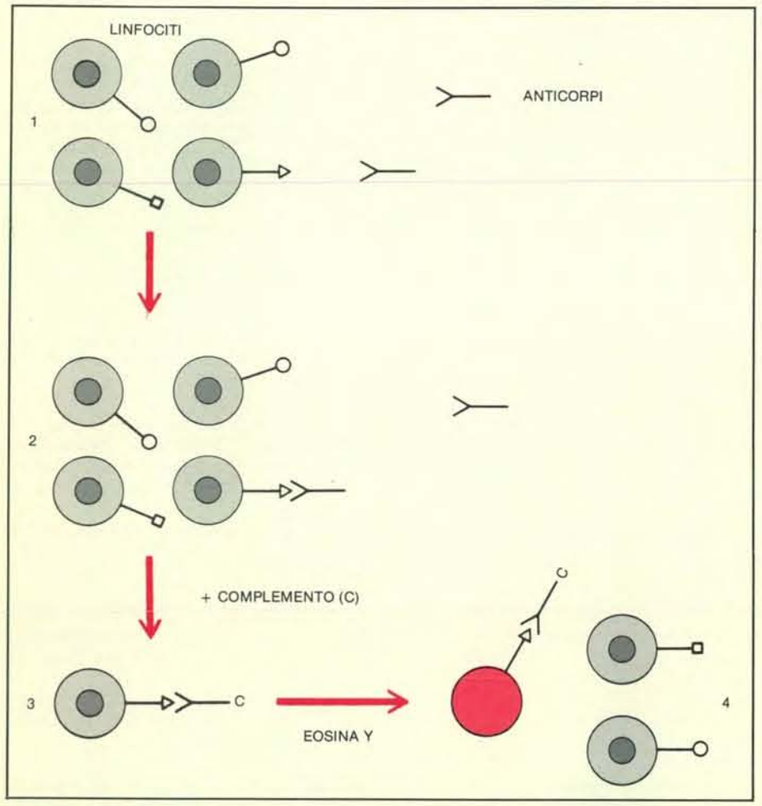
Autotrapianto e allotrapianto di cute eseguiti in due ratti di ceppo non isogenico. L'autotrapianto (*nel ratto più in alto*) è attecchito perfettamente e la cute, rasata prima dell'innesto, si è ricoperta di pelo. L'allotrapianto, coartato e necrotico, è in fase avanzata di rigetto. In basso a sinistra è visibile un allotrapianto di cute su orecchio di coniglio pochi giorni dopo l'innesto. In questa fase gli allotrapianti sembrano sempre attecchiti, ma non appena si instaura la reazione immunitaria contro gli antigeni di istocompatibilità (10-15 giorni) il lembo cade in necrosi (*a destra*).

Il maggior complesso *H-2* di istocompatibilità del topo, situato nel diciassettesimo autosoma, comprende quattro regioni dette *K, I, S* e *D;* la regione *I*, a sua volta, è suddivisa nei loci *Ir-1A, Ir-1B* e *I-C*. I principali antigeni e le funzioni delle regioni di *H-2* sono indicate dai dischetti.

ziale della parola inglese *Histocompatibility* e il numero *2* indica il locus che genera le più intense reazioni di rigetto dell'allotrapianto. Il sistema *H-2* si trova sul diciassettesimo cromosoma, in stretta concatenazione con geni responsabili di alcune malformazioni e tumori, quali quello per la coda corta o brachiuria (*Tt*) e per la crescita del pelo a ciuffi (*Tf*) e quello che in corso di leucemia timica determina la comparsa dell'antigene TL (*Thymus Leukemia*) sulle cellule linfoidi di ceppi di topi che normalmente sono sprovvisti di questo marcatore. La regione cromosomica compresa fra il locus *Tf* e quello *TL* è suddivisa in quattro subregioni, dette *K - I - S - D*.

I loci *H2-K* e *H2-D* inducono sulle cellule la formazione di quegli antigeni di istocompatibilità che si possono riconoscere per mezzo di antisieri specifici, cosicché essi sono detti antigeni SD (*Serologically Defined*). In sintesi, una cellula che veicoli un determinato antigene SD viene posta a contatto con l'antisiero, o meglio l'anticorpo, omologo. A reazione avvenuta, si aggiunge complemento che, fissandosi al complesso immune ancorato al plasmalemma, determina la morte della cellula. Le cellule uccise assumono l'eosina *Y*, al contrario di quelle viventi che non permettono a questo colorante di penetrare in esse. Usando particolari accorgimenti microscopici, sarà possibile distinguere le cellule morte, che sono uniformemente colorate in rosso opaco, da quelle vive che si presentano come dischetti intensamente luminosi. Gli elementi viventi sono quelli che mancano dall'antigene SD, corrispondente all'anticorpo contenuto nel siero utilizzato per le prove diagnostiche per cui non si è formato l'immunocomplesso capace di legare il complemento e, di conseguenza, di uccidere la cellula. Il metodo ora descritto è noto come prova della citotossicità o, più comunemente, della linfotossicità poiché abitualmente le cellule usate come bersaglio degli anticorpi sono linfociti, la cui membrana cellulare è assai ricca di antigeni di istocompatibilità.

Se due topi hanno uguale costituzione genica *H-2*, come è il caso degli animali appartenenti a uno stesso ceppo murino, possono scambiarsi qualsiasi tessuto o organo, poiché in essi non si attuerà alcuna risposta immune contro gli antigeni di istocompatibilità che sono evidentemente identici. Invece, se gli animali appartengono a ceppi diversi per il locus *H-2*, gli innesti portano antigeni di istocompatibilità differenti e sarà inevitabile la reazione da allotrapianto. Occorre però dire che oltre agli antigeni SD, le regioni *K* e *D* (e ancor più, come vedremo fra poco quella *I*), specificano anche altri determinanti che non sono definibili sierologicamente, ma per mezzo di reazioni di immunità cellulare, i quali sono detti LD (*Lymphocyte Defined*). In altri termini, gli antigeni LD non vengono riconosciuti con antisieri specifici, ma mediante l'attivazione o blastigenesi di linfociti nelle reazioni di coltura mista leucocitaria. Il fenomeno della blastigenesi ossia la trasformazione dei piccoli linfociti in grandi linfociti o immunoblasti capaci di moltiplicarsi vivacemente (i piccoli linfociti in genere non si moltiplicano), si attua quando popolazioni linfocitarie provenienti da soggetti istoincompatibili vengono mescolate e incubate a 37 °C per una settimana in atmosfera arricchita di CO_2. Quindi si aggiunge timidina triziata, si estraggono le nucleoproteine e si misura la loro radioattività. Il DNA sintetizzato dai blasti conterrà l'isotopo radioattivo e l'attività rilevata sarà proporzionale alla proliferazione linfocitaria prodottasi, ossia alla disparità degli antigeni LD delle due popolazioni di cellule esaminate. Si tratta di un fenomeno bidirezionale, nel quale entrambe le popolazioni linfocitarie tendono a blastizzare, perché si stimolano reciprocamente. La reazione può essere resa monodirezionale impedendo a una delle due popolazioni cellulari di moltiplicarsi, pretrattandola con radiazioni ionizzanti, Mitomicina C o monoiodoacetato per inibire la replicazione del DNA. I linfociti bloccati, pur essendo incapaci di blastizzare, mantengono l'attitudine a stimolare la moltiplicazione delle cellule dell'altro componente della coppia, le quali non sono state inibite artificialmente.

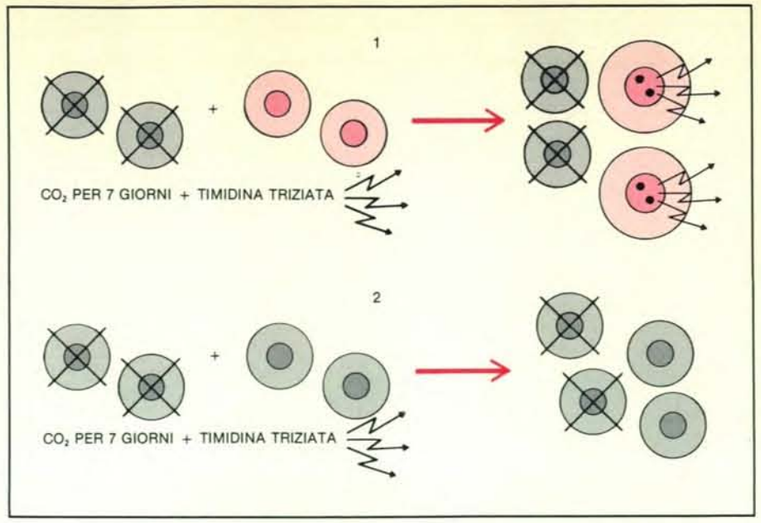


Reazione di linfotossicità. Gli anticorpi di un siero immune specifico si legano solo ai linfociti che possiedono l'antigene di istocompatibilità corrispondente (*un triangolo in figura*). Il complemento reagisce con gli anticorpi legati ai linfociti, provocandone la morte. Se si aggiunge una goccia di un colorante rosso, come l'eosina Y, le cellule uccise si distinguono facilmente al microscopio perché, a differenza di quelle vive, lasciano penetrare il colorante.

Studi condotti su particolari ceppi di topi detti congenici, i quali sono identici per gli antigeni SD ma differenti per quelli LD, hanno dimostrato che essi rigettano i trapianti più facilmente di topi congenici identici per gli antigeni LD ma non per quelli SD. Ciò è come dire che i trapianti attecchiscono più facilmente fra topi LD piuttosto che SD identici, anche se sono note alcune deroghe. Come vedremo in seguito, un comportamento simile si osserva per gli antigeni definiti dal maggior complesso di istocompatibilità umano.

La regione *I* del complesso *H-2*, che è attualmente sottoposta a intensi studi, è stata divisa in tre parti, dette *Ir-1A*, *Ir-1B* e *I-C*, dove la sigla *Ir* corrisponde a *Immune response* o risposta immune. Infatti, è stato accertato che la capacità di produrre anticorpi contro certi antigeni di struttura nota è sotto il controllo di geni situati in *Ir-1A* e *Ir-1B*. Per esempio, se topi di ceppi CBA e C57 vengono immunizzati con la stessa quantità di un polipeptide sintetico detto (T, G)-A--L, composto da tirosina, acido glutammico, alanina e lisina, si nota che i primi producono dieci volte meno anticorpi anti-(T,G)-A--L dei secondi. Se si incrociano i topi CBA con quelli C57, gli ibridi di prima generazione CBA/C57 daranno una risposta anticorpale intermedia fra quella dei genitori. Infine, i figli nati dal successivo accoppiamento di ibridi CBA/C57 × CBA/C57 si comporteranno immunologicamente come i genitori in metà dei casi, mentre un quarto sarà rappresentato da ottimi (come i C57) e un quarto da pessimi (come i CBA) produttori di anticorpi anti-(T,G)-A--L. Ciò significa che il gene che controlla la sintesi di anticorpi contro il polipeptide sintetico viene trasmesso secondo le leggi di Mendel come un carattere autosomico dominante e, soprattutto, che la capacità di produrre anticorpi è una virtù geneticamente determinata e non acquisita come si era creduto fino a poco fa.

Osservazioni simili sono state eseguite in diverse altre specie animali utilizzando sia antigeni sintetici che naturali, come per esempio l'ovoalbumina e la tireoglobulina.



Reazione leucocitaria mista. I linfociti provenienti da due individui differenti per gli antigeni LD si stimolano a vicenda a trasformarsi in blasti e a proliferare incorporando timidina triziata nel loro DNA (reazione bidirezionale). La trasformazione blastica può essere resa monodirezionale (limitata a una sola popolazione) se l'altra viene bloccata con agenti antimitotici (*cellule sbarrate*).

La regione *I-C* non sembra coinvolta nel controllo genetico dell'anticorpopoiesi, ma piuttosto nella produzione dei cosiddetti antigeni Ia (*Immune response associated antigens*), i quali sono però generati anche da *Ir-1A* e, in misura minore, *Ir-1B*. Si tratta di glicoproteine aventi peso molecolare 30 000, situate sul plasmalemma, che sono le principali responsabili del riconoscimento reciproco cellulare e, di conseguenza, dell'attivazione o blastigenesi di linfociti che si attua nella reazione da coltura mista leucocitaria, sopra ricordato.

Sempre grazie all'impiego di ceppi di topi congenici è stato dimostrato che i più potenti antigeni LD, ossia quelli capaci di dare le più forti reazioni di blastizzazione, sono codificati dalla regione *I* (e quindi identificabili con quelli Ia), mentre alle subregioni *K* e *D* spetta la produzione degli antigeni SD che inducono solo deboli fenomeni di attivazione linfocitaria.

Le stesse deduzioni valgono nel caso del fenomeno detto della «reazione del trapianto contro l'ospite» o *graft-versus-host reaction*, che è una peculiare malattia immunologica che si osserva quando i linfociti adulti vengono iniettati in topi neonati. La malattia è tanto più grave quanto maggiori sono le differenze antigeniche dettate dalla regione *I*, mentre le disparità dovute a *K* e *D* causano solo modesti fenomeni patologici. Ma quello che è ancora più interessante è che anche la suscettibilità a contrarre certe malattie, come per esempio la leucemia da virus di Gross, è correlata a geni localizzati nella regione *I* del maggior complesso *H-2*; si tratta di un fenomeno che si ripete con modalità analoghe nell'uomo. Un accenno merita la regione *S* del complesso *H-2* che sembra controllare la produzione quantitativa e qualitativa del quarto componente del complemento e di una peculiare β-globulina sierica.

Come si vede, il maggior complesso di istocompatibilità murino *H-2* è il supervisore di tutte le manifestazioni immunitarie del topo, che comprendono la capacità di distinguere gli antigeni propri da quelli estranei, di rigettare i trapianti e di produrre anticorpi umorali e persino di modulare la sintesi di fattori del complemento, ossia di un sistema di proteine ancillare a quello immunologico.

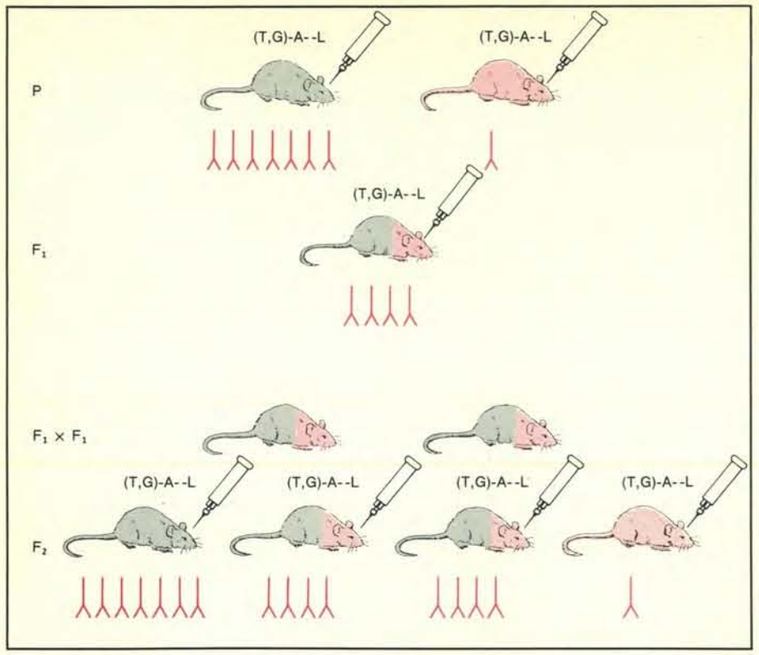
Nell'uomo il maggior complesso di istocompatibilità è contrassegnato dalla sigla *HLA*, dove le lettere *H* e *L* sono le iniziali delle parole *Human* e *Leukocyte* e *A* indicava originariamente il primo locus geneticamente definito per gli antigeni leucocitari. In questi ultimi anni le conoscenze sugli antigeni determinati dal sistema *HLA* e sui loro rapporti genetici hanno subito un eccezionale incremento grazie a una serie di *International Histocompatibility Workshops*, dopo ognuno dei quali un Comitato per la nomenclatura, patrocinato dalla Organizzazione mondiale della sanità, ha valutato l'insieme dei dati raccolti e proposto una terminologia unificata per gli antigeni di istocompatibilità di volta in volta identificati. Nel 1967 ci si rese conto che i loci *HLA* erano due e non uno, posti assai vicini sullo stesso cromosoma, cosicché due antigeni dettati dal cromosoma venivano quasi sempre ereditati insieme. I due loci, che vennero definiti primo o *LA*, perché includeva alcuni antigeni allora detti LA-1, LA-2 etc., e secondo o *4*, perché includeva gli antigeni 4a e 4b, codificavano complessivamente sei antigeni di istocompatibilità ben riconoscibili. Il Workshop del 1970 aumentò da 6 a 20 il numero degli antigeni che venivano identificati dalla sigla HLA (allora scritta HL-A), seguita da un numero arabo, se accettati definitivamente e da quella W (*Workshop*), seguita da un numero arabo, se accettati provvisoriamente dal Comitato per la nomenclatura.

Nel 1972 il numero degli antigeni salì a 31; ci si rese conto che alcuni di quelli precedentemente considerati una singola entità erano in realtà composti da diversi sottotipi e si stabilì l'esistenza di un terzo locus che venne denominato *AJ*. Fu inoltre accertato che la blastigenesi dei linfociti umani nelle reazioni da cultura mista leucocitaria era causata da differenze antigeniche determinate da un quarto locus della regione *HLA*, che venne detto *MLR* (*Mixed Leukocyte Reaction*). Poiché gli antigeni determinati da *LA*, *4* e *AJ* sono identificabili sierologicamente e quelli determinati da *MLR* per mezzo dell'attivazione linfocitaria, si pensò di ribattezzare i primi tre loci, SD_1, SD_2, e SD_3 (*Serologically Defined* $_{1\ 2\ 3}$) e il quarto LD_1 (*Lymphocyte Defined*$_1$).

Dopo il Workshop del 1975, che ha portato il numero degli antigeni a 51, il Comitato per la nomenclatura ha proposto una terminologia normalizzata per la quale i loci SD_1, SD_2, SD_3 e LD_1 vengono detti *HLA-A*, *HLA-B*, *HLA-C*, e *HLA-D*.

Questi loci sono situati nel sesto autosoma e sono strettamente concatenati con i loci per la adenosindeaminasi (*ADA*), la fosfoglicomutasi (PGM_3), il fattore B della via alternativa del complemento (*GBG* = *Glycin-rich-β-glycopeptide*) e per il gruppo ematico *Chido*.

Se si confrontano fra loro i maggiori complessi di istocompatibilità *HLA* e *H-2*, la differenza più rilevante è che nell'uomo la regione genetica che regola la blastizzazione nelle reazioni da cultura mista leucocitaria, e verosimilmente altre risposte immuni analoghe a quelle controllate dalla regione *I* del topo, si trova



La capacità di produrre anticorpi è trasmessa come carattere dominante, come dimostra questo esperimento che riguarda la sintesi di anticorpi verso il polipeptide (T,G)-A- -L, nei discendenti del ceppo C57, buon produttore di anticorpi, e CBA, cattivo produttore di anticorpi.

all'esterno delle regioni *HLA-A, HLA-B* e *HLA-C* che codificano gli antigeni da trapianto sierologicamente definiti, contrariamente a quanto avviene nel topo, specie nella quale si verifica la situazione diametralmente opposta.

Sebbene i loci *HLA-A, B, C, D* possano codificare 10, 10, 5 e 6 differenti antigeni, avviene che in ogni singolo individuo non sia presente sulle cellule più di un antigene di istocompatibilità per locus.

Poiché nei vertebrati i cromosomi sono disposti a coppie formate da un elemento di origine materna e uno di origine paterna, se ne dedurrà che sulle cellule ci sono due antigeni differenti prodotti dallo stesso locus se i due cromosomi sono uguali (individuo eterozigote) e un solo antigene se i due cromosomi sono uguali (individuo omozigote). Per esempio, se i cromosomi della madre e del padre portano l'uno l'allele *HLA-A1* e l'altro l'allele *HLA-A3*, le cellule veicoleranno gli antigeni HLA-A1 e HLA-A3, mentre se entrambi portano l'allele *HLA-A1*, le cellule avranno il corredo antigenico HLA-A1, sebbene in doppia dose HLA-A1, HLA-A1. I loci *HLA-A, B, C, D* sono molto vicini fra di loro sul cromosoma e i determinanti antigenici specificati da essi vengono ereditati in un unico blocco, che viene detto aplotipo. Per esempio, un padre che abbia gli aplotipi *HLA-A1, B8, Cw1, Dw2 / HLA-A3, B7, Cw2, Dw6* trasmetterà sempre ai figli, salvo rarissime eccezioni, l'aplotipo *HLA-A1, B8, Cw1, Dw2* oppure quello *HLA-A3, B7, Cw2, Dw6* e in modo analogo si comporterà la madre. È consuetudine definire gli aplotipi paterni con le lettere A e B e quelli materni con quelle C e D. Quando si debba eseguire un trapianto d'organo, come quello di rene, e il donatore è un famigliare dell'accettore, è molto importante stabilire se gli aplotipi della coppia donatore-accettore sono identici o differenti. Infatti, nel caso di una coppia *HLA*-identica, l'operazione di trapianto sarà quasi certamente coronata da successo, in termini di sopravvivenza dell'organo trapiantato, mentre ciò non si verifica se la coppia ha un solo aplotipo in comune (*HLA*-aploidentica) o, peggio, nessuno (*HLA*-differente). Per stabilire l'ereditarietà degli aplotipi occorre tipizzare i linfociti periferici dell'accettore, dei suoi fratelli e sorelle e dei genitori, utilizzando il metodo della linfocitototossicità precedentemente descritto. Purtroppo solo pochi laboratori sono in grado di tipizzare i determinanti antigenici specificati dai loci *HLA-C* e *HLA-D* (questi ultimi devono essere studiati con la tecnica della blastizzazione *in vitro*) e nella maggior parte dei casi ci si deve limitare all'analisi di quelli codificati da *HLA-A* e *HLA-B*.

Però, poiché l'aplotipo viene ereditato in blocco si presume che due fratelli identici per i loci *HLA-A* e *HLA-B* lo siano anche per *HLA-C* e soprattutto *HLA-D* che, come sappiamo, è il locus le cui differenze sono maggiormente responsabili del rigetto dei trapianti. Occorre tenere presente che a uguali fenotipi non corrispondono necessariamente uguali aplotipi. Per esempio: 1) due antigeni simili ma non identici, come lo HLA-3 e HLA-11 possono dare reazioni crociate con gli anticorpi, onde un soggetto HLA-3 può essere ritenuto HLA-11 e viceversa, 2) alcuni antigeni possono sfuggire all'identificazione con i sieri attualmente disponibili e 3) se i genitori hanno un antigene in comune, i figli possono avere lo stesso fenotipo che però è generato da differenti aplotipi. Se ne deduce che nel caso di un trapianto di rene occorre tipizzare il più completamente possibile l'accettore e i famigliari candidati alla donazione e quindi analizzare attentamente l'ereditarietà degli aplotipi. Per esempio, supponiamo che in una famiglia vi siano quattro fratelli con i seguenti fenotipi: A1, A3, B5, B8, - A1, A9, B7, B8 - A2, A3, B5, B12 - A2, A9, B7, B12, che il padre sia A1, A2, B8, B12 e la madre A3, A9, B5, B7. Si potrà stabilire che gli aplotipi del padre sono (A) = *A1, B8* e (B) = *A2, B12* e quelli della madre (C) = *A3, B5* e (D) = *A9, B7*.

Infatti nel primo figlio gli antigeni A1, B8 devono provenire dal padre perché mancano nella madre e ciò indica appunto che un aplotipo del padre è *A1, B8* e l'altro, necessariamente, *A2, B12*. Lo stesso figlio mostra che gli aplotipi della madre sono *A3, B5,* e *A9, B7* per ragioni analoghe alle precedenti. I risultati della tipizzazione degli altri tre figli, nei quali gli aplotipi sono presenti in tutte le possibili combinazioni, confermano il dato. Gli aplotipi dei quattro figli sopra indicati sono, rispettivamente, AC, AD, BC e BD. Come si nota, vi sono coppie *HLA*-aploidentiche (AC-AD; BC-BD; AC-BC; AD-BD) e *HLA*-differenti (AC-BD; AD-BC) ma nessuna *HLA*-identica. Però, se questa famiglia avesse un quinto figlio, esso dovrebbe essere *HLA*-identico con un fratello, poiché non vi sono altre combinazioni aplotipiche possibili oltre a quelle indicate.

D'altra parte è evidente che anche in una famiglia con due soli figli questi possono essere *HLA*-identici, poiché la trasmissione ereditaria degli aplotipi avviene del tutto casualmente e che i genitori sono sempre *HLA*-aploidentici con ognuno dei loro figli.

Nella maggior parte dei trapianti renali sinora eseguiti, qualche decina di migliaia, il rene innestato è di origine cadaverica. In questo caso la sopravvivenza dell'organo è abitualmente minore di quella osservata fra fratelli *HLA*-identici e si nutrono molte perplessità sul valore della tipizzazione linfocitaria. Infatti i ricercatori europei affermano che la durata della sopravvivenza dell'innesto è direttamente proporzionale alla somiglianza degli antigeni HLA del donatore con quelli dell'accettore, mentre i nordamericani negano la significatività della correlazione. In alcuni centri si sostiene che la sopravvivenza del rene cadaverico è statisticamente aumentata solo se l'accettore manca di anticorpi contro gli antigeni HLA del donatore e se vi è identità almeno al locus *HLA-B*. Occorre perciò eseguire preliminarmente a ogni intervento chirurgico una reazione crociata fra siero dell'accettore e linfociti del donatore per escludere la prima evenienza la cui incidenza è tanto minore quanto migliore è stato l'appaiamento fra gli antigeni HLA della coppia donatore-accettore.

I migliori risultati che si osservano quando vi è identità al locus *HLA-B* piuttosto che allo *HLA-A* è verosimilmente attribuibile al fatto che il locus *HLA-B* è il più vicino a quello *HLA-D*, responsabile della blastigenesi linfocitaria e, in definitiva, della reazione di rigetto. In altre parole, due individui non consanguinei che abbiano gli stessi antigeni HLA-B sono probabilmente identici anche per quelli dettati dal locus *HLA-D*. Ciò è più verosimile per la popolazione europea che è generalmente più omogenea (più *inbred*) di quella nordamericana e proprio in questa osservazione giace la spiegazione delle differenti valutazioni sull'importanza della tipizzazione *HLA* espresse da parte dei ricercatori dei due continenti.

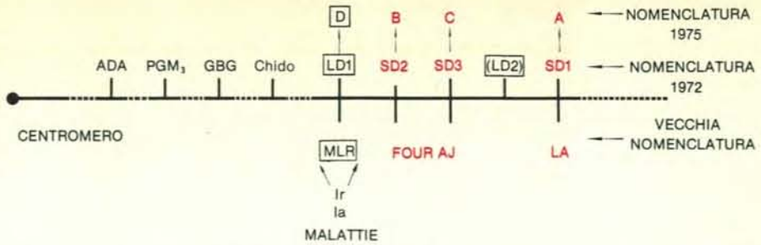
Recentemente alcuni immunologi inglesi hanno addirittura proposto di tipizzare solo per due particolari antigeni sierologicamente determinati detti 4a e 4b la cui mappatura genetica è ancora incerta, anche se si ha ragione di assegnarli al locus *HLA-B*, con notevole risparmio di tempo e denaro e con risultati uguali a quelli che si ottengono con le indaginose prove di linfocitotossicità per tutti gli antigeni determinati da *HLA-A* e *HLA-B*. La questione della tipizzazione verrebbe superata se i trapianti venissero scambiati solo fra individui i cui linfociti non si stimolano a vicenda nelle reazioni da coltura mista linfocitaria. Purtroppo, questa prova richiede almeno sette giorni di tempo e non può evidentemente essere eseguita nel caso di trapianto da donatore cadaverico perché il rene isolato sopravvive poche ore e deve essere innestato il più presto possibile. La soluzione del problema richiede: 1) di abbreviare i tempi delle reazioni di coltura mista leucocitaria, 2) di allungare i tempi di conservazione degli organi isolati, 3) di tipizzare i determinati HLA-D con anticorpi sierici come sembra si possa fare a patto che le cellule-bersaglio siano costituite da linfociti *B* e 4) di usare cellule-bersaglio già «innescate» e perciò capaci di blastizzare con maggiore rapidità.

Le osservazioni che si riferiscono al rene si applicano anche ai trapianti di altri organi, come cuore, fegato, polmoni ecc... Però nessun centro ha sinora eseguito tanti trapianti di questo genere quanti sono necessari a valutare l'importanza della tipizzazione *HLA*. Quello che è certo è che nessun organo dovrebbe venir innestato prima che le reazioni crociate fra linfociti del donatore e siero dell'accettore abbiano dimostrato che il secondo è privo di anticorpi contro gli antigeni di istocompatibilità del primo.

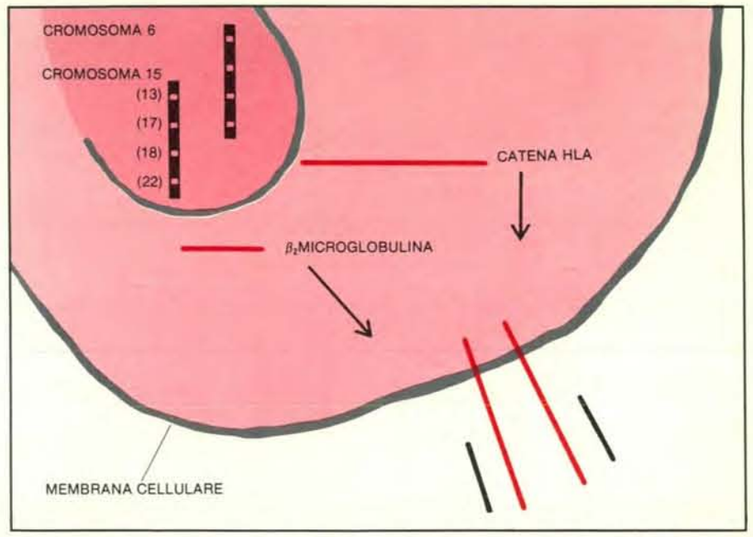
Il maggior complesso *HLA* esercita le altre funzioni oltre quella della sintesi degli antigeni da trapianto. Come si è già detto, esso controlla la capacità individuale di produrre anticorpi e la suscettibilità a contrarre certe malattie. Per quanto riguarda il primo aspetto, le informazioni sui geni delle risposte immuni dell'uomo sono assai più scarse di quelle afferenti ai geni *Ir* del sistema *H-2* del topo. Tuttavia è lecito ritenere che geni *Ir* esistano anche nell'uomo e che siano localizzati presso il locus *HLA-D*. Per esempio, i soggetti che dispongono dell'antigene HLA-A3 sono capaci di produrre grandi quantità di anticorpi contro quello HLA-A2, mentre ciò non si verifica in quelli privi di HLA-A3. Per questa ragione, volontari umani HLA-3 vengono spesso immunizzati con cellule HLA-A2, per produrre antisieri specifici contro quest'ultimo antigene.

Si è anche osservato che i linfociti di soggetti HLA-B5 blastizzano facilmente se coltivati in presenza di antigeni streptococcici al contrario di quelli di individui non HLA-B5, facendo pensare che vi è un'associazione significativa fra elevata risposta e HLA-B5. È stata rilevata anche una netta correlazione fra certi aplotipi *HLA* e la risposta immune contro l'antigene *E* del polline di *ragweed*, una pianta anemogama del genere *Ambrosia* che è responsabile di buona parte dei casi di «febbre da fieno» che si verificano nell'America settentrionale. Per esempio, in una numerosa famiglia si osservò che 20 delle 26 persone affette dalla malattia veicolavano sempre allo stesso aplotipo, che era costantemente assente nei consanguinei sani, cosicché si pensa che tale aplotipo è associato con un gene *Ir* che controlla sia la sintesi di *IgE* che di *IgG* contro l'antigene *E*.

Naturalmente, queste ed altre osservazioni non provano ancora definitivamente l'esistenza di geni *Ir* nell'uomo, ma occorre ricordare che le ricerche in campo umano soffrono di gravi limitazioni dalle quali sono esenti quelle in campo murino. Infatti non è possibile immunizzare l'uomo con quella vasta gamma di antigeni sintetici e naturali utilizzata nel topo. In quest'ultimo è stata localizzata la regione *I* usando sieri contro gli antigeni Ia, che non sono disponibili per



Nel maggior complesso *HLA* di istocompatibilità dell'uomo, situato nel sesto autosoma, si distinguono tre loci *SD*, ossia *A, B,* e *C*, e un locus *LD*, denominato *D*; l'esistenza di un secondo locus *LD2* è ancora da accertare. La posizione dei loci umani, corrispondenti a quelli murini *Ir* (risposte immuni), *I-C* (antigeni Ia) e di quelli che controllerebbero certe malattie è sommariamente indicata.



La specificità immunologica HLA spetta alle catene pesanti, che sono codificate dai geni *SD* del maggior complesso d'istocompatibilità. Le catene leggere, o β2-microglobuline, hanno una struttura primaria omologa a quella dei domini costanti delle IgG e sono forse codificate dall'autosoma 15.

MALATTIE	ANTIGENE	RISCHIO RELATIVO
SPONDILITE ANCHILOSANTE	B27	120,9
SINDROME DI REITER	B27	40,3
UVEITE ACUTA ANTERIORE	B27	30,7
DIABETE MELLITO INSULINO-DIPENDENTE	B8	2,1
	Bw15	3,0
	Dw3	4,5
	Dw4	3,7
* RISCHIO NORMALE = 1,0		

Nei possessori di certi antigeni del sistema HLA si manifestano malattie con frequenza superiore alla restante popolazione. Questa osservazione può essere utile per la loro diagnosi e profilassi.

l'uomo (a eccezione di quelli del recentissimo 7th International Histocompatibility Workshop), e infine gli antisieri che individuano gli antigeni di istocompatibilità umani sono meno specifici dei reagenti murini. Spesso un siero reagisce crociatamente con diversi antigeni HLA, rendendo difficile l'accertamento dell'aplotipo del soggetto in esame.

Certe malattie si manifestano in maniera significativamente più frequente nei soggetti possessori di determinati antigeni di istocompatibilità rispetto a quelli che ne sono privi. Per esempio, chi porta l'antigene HLA-B27 ha probabilità di contrarre la spondilite anchilosante, la sindrome di Reiter (artrite, uretrite, congiuntivite) o l'uveite anteriore acuta con frequenza 121, 40 e 31 volte superiore a colui che manca di tale fattore. Il diabete mellito insulino-dipendente è associato con gli antigeni HLA-B8, HLA-Bw15, HLA-Dw3 e HLA-Dw4 con frequenza 2, 3, 5 e 4 volte superiore alla norma. Tecnicamente il valore numerico che indica la maggior frequenza è detto «rischio relativo» e si contrappone al «rischio normale», che è uguale a 1, di incorrere in una determinata forma morbosa.

In questo campo esistono numerosi studi con risultati spesso contradditori. Infatti, la frequenza nella popolazione di certi aplotipi che, ricordiamo, sono la combinazione di un antigene dettato dal locus *HLA-A* con uno del locus *HLA-B* (a parte gli *HLA-C* e *HLA-D* che non vengono inclusi in questa discussione, come precedentemente detto) è maggiore di quanto si possa prevedere in base al calcolo statistico delle combinazioni. Per esempio, le combinazioni HLA-A1; HLA-B8 e HLA-A3; HLA-B8 predominano nettamente sulle altre fra i caucasici e se non si introducessero nel calcolo del rischio relativo gli opportuni correttivi statistici, si incorrerebbe nel grossolano errore di attribuire a tali aplotipi quasi tutte le malattie dei bianchi.

È stato ipotizzato che certi antigeni HLA possono fungere da recettori per i virus e facilitare l'insorgere di particolari malattie. Un'interazione di questo tipo fra gli aplotipi dei diabetici insulino-dipendenti e il virus *Coxsackie B4* rappresenterebbe il primo passo del danneggiamento delle cellule degli isolotti pancreatici di Langerhans che si traduce poi nel diabete giovanile insulino-dipendente.

Se queste osservazioni venissero estese e provate in modo inconfutabile, se ne dedurrebbe che tutte le malattie, comprese quelle infettive, sono sotto il controllo del genotipo e che ognuno di noi ha un certo destino nosologico. È però ragionevole e auspicabile pensare che i fattori ereditari siano una causa necessaria, ma non sufficiente, di malattia, la cui attuazione richiederebbe l'intervento di concause ambientali precipitanti.

Da questo insieme di informazioni emerge che fra le funzioni attribuite al maggior complesso di istocompatibilità, quella di maggior spicco è il rigetto dell'allotrapianto. Indipendentemente dal fatto che si tratta anche del fenomeno meglio studiato per le sue applicazioni nel campo della chirurgia sostitutiva umana, viene spontaneo chiedersi quale finalismo evoluzionistico ha portato alla creazione nella specie uomo e negli altri vertebrati finora esaminati, di un sistema così complesso e preciso per ostacolare l'attecchimento dei tessuti estranei, dato che nessun tipo di trapianto si attua spontaneamente in natura, con la sola eccezione della gravidanza. Infatti il feto che cresce nella cavità uterina è un allotrapianto *HLA*-aploidentico nei confronti della madre, poiché metà dei suoi antigeni di istocompatibilità sono di derivazione paterna. Tuttavia esso viene tollerato per nove mesi grazie a meccanismi protettivi la cui esatta natura ci sfugge ancora e rappresenta proprio l'eccezione che conferma la regola.

È stato ipotizzato che il ruolo predominante del maggior complesso di istocompatibilità è quello di rigettare le cellule maligne che si formano continuamente nell'organismo. Queste cellule, che veicolano antigeni di istocompatibilità tumorali detti TSTA (*Tumor Specific Transplantation Antigens*) differenti da quelli normali, vengono identificate come estranee dal sistema immunocompetente del loro involontario ospite ed eliminate grazie alla reazione da allotrapianto. Se così non fosse, certamente l'incidenza dei tumori maligni sarebbe assai superiore a quella attuale, poiché ogni giorno miriadi di cellule nascono e muoiono nel corpo di ogni essere vivente e molte di esse mutano assumendo caratteristiche tumorali.

In certi individui però, un sottile difetto della capacità di riconoscere gli antigeni tumorali o di quella di organizzare rapidamente un'efficiente reazione anticorpale permette al tumore di crescere in breve tempo fino a raggiungere dimensioni tali da soverchiare le capacità difensive del sistema immunitario.

Esistono molte osservazioni a conferma della teoria della «sorveglianza immunologica.» Per esempio, individui geneticamente immunodeficienti sviluppano neoplasie maligne con frequenza superiore alla media e lo stesso dicasi per i soggetti volutamente immunodepressi con mezzi farmacologici, biologici e fisici come è il caso degli accettori di trapianti di rene, nei quali l'incidenza dei tumori è assai più elevata di quella della popolazione normale di controllo. Inoltre, certi topi che sono privi di timo alla nascita (topi nudi), che è come dire che sono incapaci di produrre reazioni da allotrapianto, accettano normalmente innesti di cellule cancerose che vengono rigettate dagli animali normali.

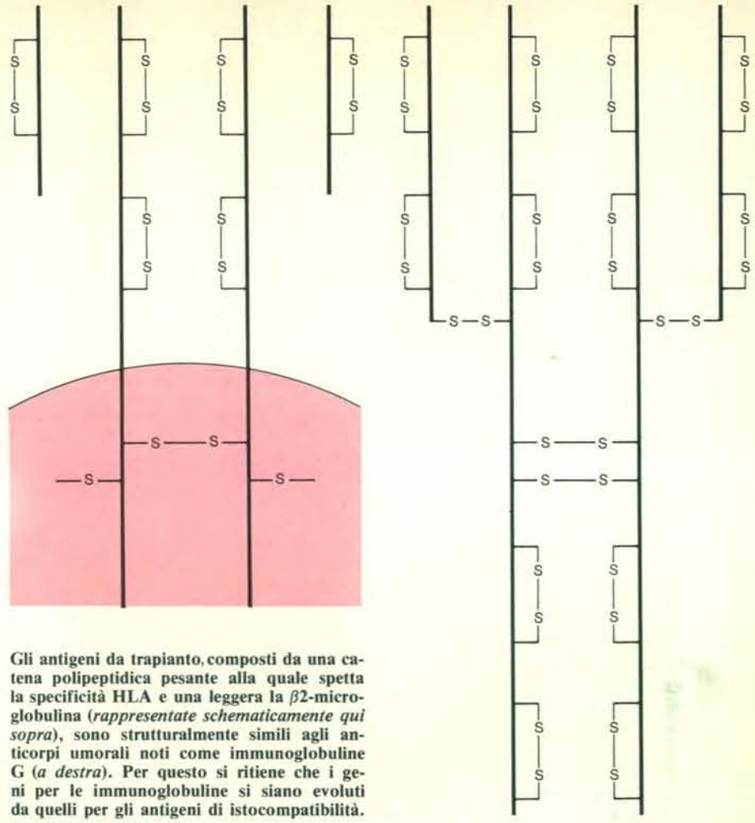
In conclusione, il maggior complesso di istocompatibilità *HLA* controlla la produzione di alcune strutture, come gli antigeni da trapianto, necessarie alla manifestazione della propria identità, e altre, come gli anticorpi e i recettori cellulari, destinati a riconoscere le strutture estranee che possono essere sia antigeni da trapianto normali e tumorali che germi patogeni o molecole innocue ma dotate di attività antigenica. È degno di nota il fatto che esistano strette analogie morfologiche fra gli antigeni di istocompatibilità e gli anticorpi umorali, che è come dire fra le molecole che devono farsi riconoscere e quelle che devono riconoscere.

Infatti gli antigeni HLA sono costituiti da quattro catene peptidiche, due di peso molecolare 33 000 e due 12 000, collegate fra loro da legami non covalenti e innicchiate nella membrana cellulare. La specificità sierologica HLA spetta alle catene pesanti mentre quelle leggere sono state identificate come β_2-microglobuline o β_2m ossia proteine originariamente isolate dalle urine di pazienti tubolopatici, le quali sono presenti in elevata quantità anche sulle membrane dei linfociti e di altre cellule. Anche gli anticorpi sono costituiti da quattro catene polipeptidiche, due di peso molecolare circa 50 000 e due circa 25 000, che però sono collegate fra loro da legami covalenti. La differenza di peso molecolare fra gli antigeni HLA e gli anticorpi potrebbe essere attribuita alla mancanza, nei primi, delle cosiddette regioni variabili, che sono invece parte del sito anticorpale dei secondi. Però è verosimile che tali regioni esistano ma che vengano rimosse durante i procedimenti tecnici necessari per l'isolamento e la purificazione degli antigeni HLA. Candidati alla funzione di regione variabile di questi antigeni potrebbero essere i prodotti dei geni *Ir* e *Ia*, i quali nel maggior complesso di istocompatibilità sono strettamente concatenati ai geni strutturali per gli antigeni da trapianto.

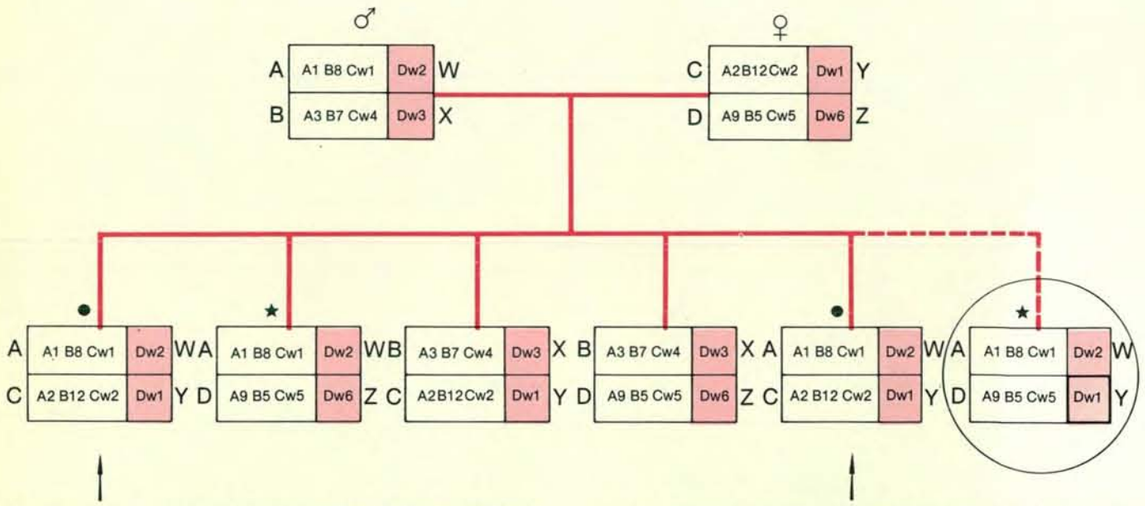
La parentela fra antigeni HLA e anticorpi va ben oltre alla semplice somiglianza morfologica poiché è stato dimostrato che la struttura amminoacidica primaria delle β_2m è molto simile a quella delle regioni costanti delle catene pesanti e leggere delle *IgG*, come se si trattasse di molecole originate da un unico o da pochi geni ancestrali.

Questi geni primordiali che condizionano l'evoluzione del sistema di riconoscimento immunologico dei vertebrati sembrano essere imparentati, almeno nel topo, con quelli del sistema multiallelico *Tt*, il quale si trova strettamente concatenato al maggior complesso *H-2* nel diciassettesimo cromosoma. I geni *Tt* sono una serie di omozigoti letali che influenzano molti passi della primitiva differenziazione embrionale e in particolare lo sviluppo della notocorda, dello scheletro assile e del tubo neurale determinando degli antigeni di superficie che si manifestano sulle cellule embrionali e anche sugli spermatozoi. Vi sono molte analogie fra i sistemi *Tt* e *H-2* che sono entrambi multigenici e che controllano numerosi fenomeni di identificazione cellulare mediati da proteine di superficie. Molti autori ritengono che il sistema *Tt* sia responsabile del riconoscimento cellulare durante le più precoci fasi dell'ontogenesi mentre il maggior complesso di istocompatibilità lo sarebbe nella tarda embriogenesi e nella vita adulta; esso inoltre si sarebbe ulteriormente evoluto fino ad assumersi l'onere di identificare gli antigeni estranei e di sintetizzare i componenti del sistema complementare, che è accessorio a quello immunologico.

Quindi la sequenza evolutiva comincerebbe con i geni che determinano le proteine di superficie necessarie per lo sviluppo normale della notocorda, dello scheletro assile e del tubo neurale che sono i processi cruciali nell'embriogenesi dei vertebrati. In seguito a fenomeni di duplicazione e diversificazione tali geni primordiali genererebbero quelli del maggior complesso di istocompatibilità e da questi ultimi si svilupperebbero, con analogo meccanismo, quelli per le immunoglobuline. In tale modo verrebbero concentrate nel maggior complesso di istocompatibilità tutte quelle funzioni di riconoscimento immunologico che sono indispensabili per la sopravvivenza e l'adattamento ambientale della specie e, addirittura, del singolo individuo.



Gli antigeni da trapianto, composti da una catena polipeptidica pesante alla quale spetta la specificità HLA e una leggera la β2-microglobulina (*rappresentate schematicamente qui sopra*), sono strutturalmente simili agli anticorpi umorali noti come immunoglobuline G (*a destra*). Per questo si ritiene che i geni per le immunoglobuline si siano evoluti da quelli per gli antigeni di istocompatibilità.



Ereditarietà degli antigeni HLA di tipo SD (A, B, C, D) e LD (W, X, Y, Z). Solo 4 combinazioni di aplotipi dei genitori sono possibili nei figli e ogni figlio ha in comune un aplotipo con ogni genitore purché non sia avvenuto un *crossing-over* nel genoma parenterale. Nell'esempio, il primo e il quinto figlio sono identici sia per gli antigeni SD che LD, mentre il secondo e il sesto sono SD identici, LD differenti.

L'atmosfera di Marte

Nonostante la sua densità sia inferiore a un centesimo di quella terrestre, ha nubi e venti analoghi ed è l'agente principale delle modificazioni della superficie del pianeta

di Conway B. Leovy

La serie coronata da successo di missioni spaziali verso Marte, iniziata nel 1964 e culminata l'anno scorso con il riuscito atterraggio dei due moduli di atterraggio (*lander*) dei Viking, ci ha permesso di conoscere quel pianeta quasi quanto la Luna (si veda l'articolo *Marte visto dai Viking* di M. Fulchignoni e M. Coradini, in «Le Scienze», n. 100, dicembre 1976). A differenza della Luna, la cui storia sembra sia finita sostanzialmente uno o due miliardi di anni fa, Marte sta ancora evolvendosi e modificandosi. Su Marte, come sulla Terra, la causa di mutamenti diffusa più estesamente è l'atmosfera del pianeta, prodotta a sua volta dalla liberazione in superficie dei costituenti iniziali del pianeta, cominciata fin dalla condensazione dalla nube primordiale di polvere e gas che diede origine al sistema solare 4,6 miliardi di anni fa.

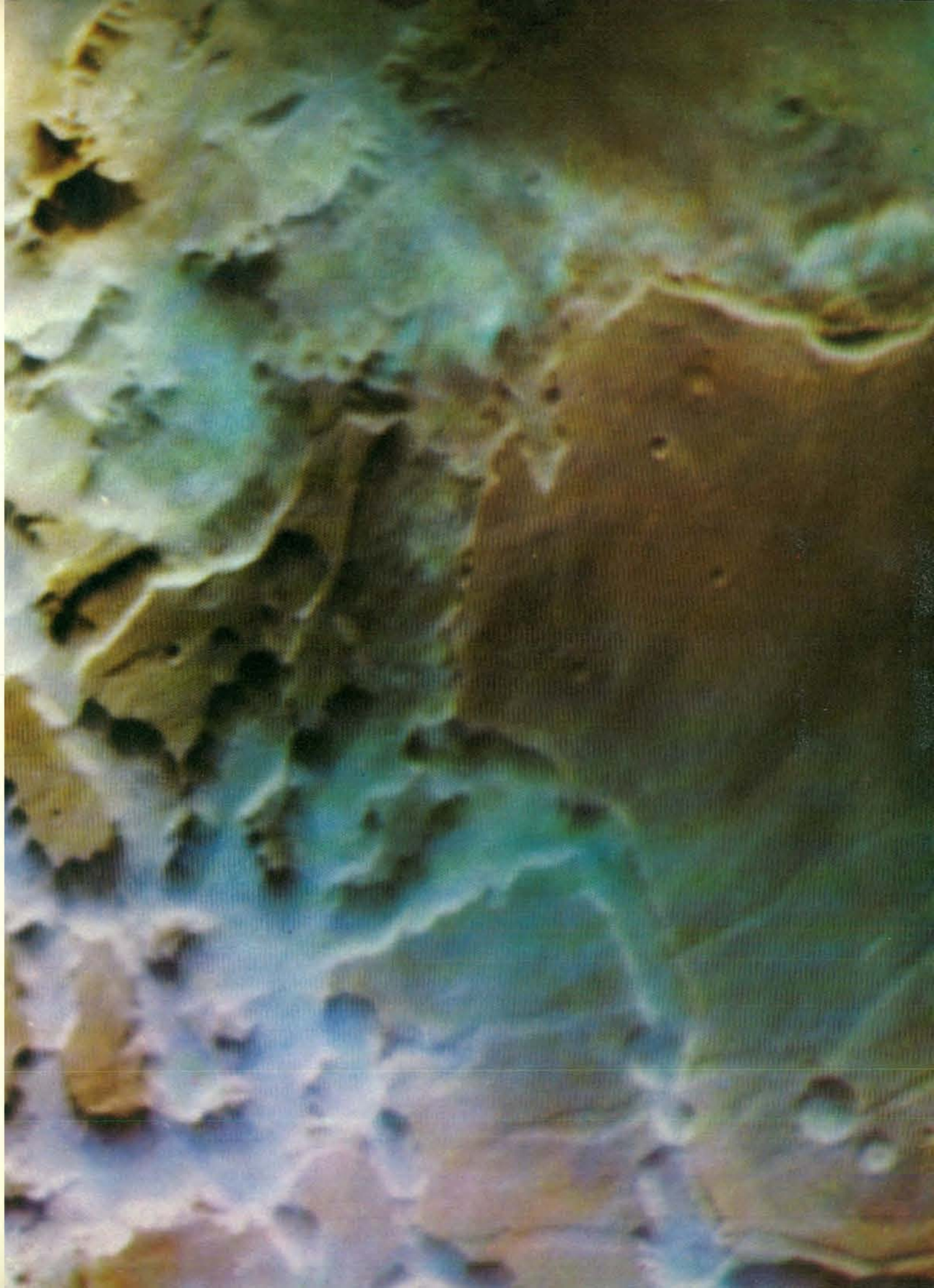
Qualche informazione sull'atmosfera di Marte era stata dedotta dalle osservazioni telescopiche prima dei voli spaziali, ma tali informazioni non erano verificabili e si prestavano a interpretazioni errate. Con i dati molto più accurati forniti dai veicoli spaziali che sono passati in prossimità di Marte, che gli hanno orbitato intorno e che infine sono atterrati su di esso, conosciamo ora una seconda atmosfera planetaria in modo sufficientemente dettagliato da permettere di verificare la nostra comprensione dell'evoluzione e dei processi dell'atmosfera basata finora sull'unico esempio disponibile: l'atmosfera terrestre. Tranne che per la sua composizione chiaramente diversa, l'atmosfera di Marte si comporta sostanzialmente come una versione rarefatta della nostra. Trasporta acqua, genera nubi e i suoi venti sembrano soffiare secondo schemi giornalieri e stagionali. A seguito delle variazioni stagionali dell'energia ricevuta dal Sole si originano tempeste di polvere localizzate, che talvolta si intensificano fino a interessare tutto quanto il pianeta. Queste tempeste di polvere su scala globale sono un fenomeno che per il momento può considerarsi unicamente marziano.

A causa degli agenti atmosferici la roccia cristallina primitiva su Marte si è sgretolata, sminuzzata in frammenti, si è ossidata e si è combinata chimicamente con l'acqua producendo quei minerali rossastri caratteristici, così ben evidenti nelle immagini trasmesse dai *lander* dei Viking. I meccanismi degli agenti atmosferici operanti su Marte sono chiaramente diversi dai processi atmosferici dominanti sulla Terra, che dipendono dall'acqua allo stato liquido. Dovunque su Marte sono presenti segni evidenti di erosione e sedimentazione eolica. Per esempio i due luoghi di atterraggio dei Viking, rispettivamente la Chryse Planitia e l'Utopia Planitia, sembra siano stati battuti fortemente dai venti. Strati profondi di sedimenti portati dai venti si sono accumulati nelle regioni polari di Marte, e campi di dune più estesi di qualsiasi analoga formazione terrestre sono stati fotografati nella regione presso il polo nord.

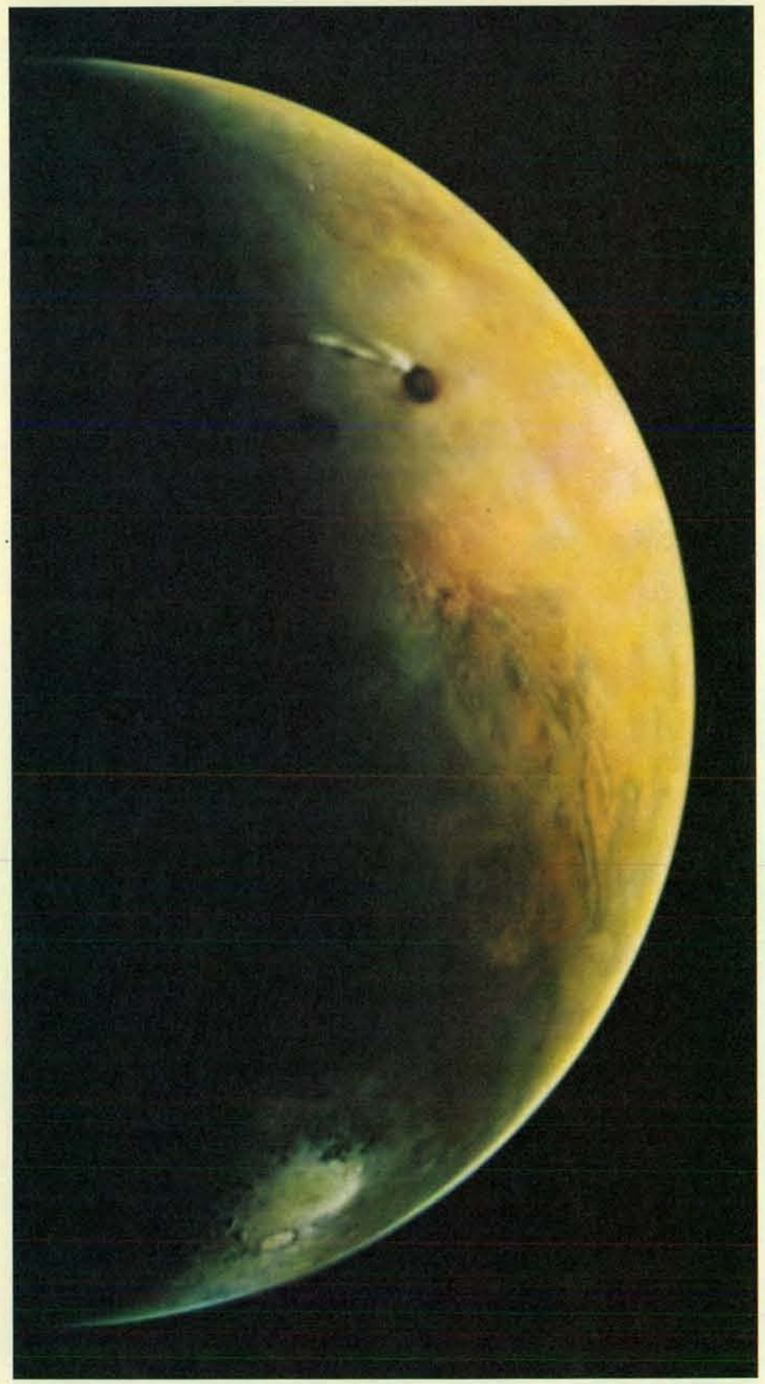
A causa della pressione atmosferica ridotta e della temperatura generalmente bassa le sole forme stabili dell'acqua trovate su Marte sono quelle di vapore acqueo e di ghiaccio. Nonostante l'attuale assenza di acqua in fase liquida, le fotografie scattate dai veicoli in orbita mostrano chiaramente che la superficie del pianeta è stata soggetta a erosione da fluidi su larga scala. Tali fluidi furono molto probabilmente acqua. È possibile che ci sia stato uno scorrere di acqua sulla superficie prodotto dalla fusione o dalla vaporizzazione di masse di ghiaccio appena sotto la superficie. È pure possibile che in passato il clima sul pianeta Marte fosse più mite e l'atmosfera abbastanza densa e umida da produrre cadute di pioggia.

Le caratteristiche più rilevanti di una atmosfera planetaria sono la sua composizione e la sua massa oppure, equivalentemente, la sua pressione superficiale. Ancora nel 1963, l'anno prima del lancio di *Mariner 4* per il suo riuscito viaggio verso Marte, si credeva, sulla base delle osservazioni telescopiche, che la pressione atmosferica presente su Marte alla superficie fosse di circa 85 millibar, un valore cioè un po' minore del 10 per cento della pressione superficiale media sulla Terra, che è di 1013 millibar. Si pensava inoltre che il gas più abbondante nell'atmosfera di Marte fosse l'azoto. L'anidride carbonica e l'acqua erano state identificate spettroscopicamente, ma si pensava che entrambe fossero componenti minori, dato che nessuna delle due appariva sufficientemente abbondante da contribuire significativamente al valore totale assunto di 85 millibar. L'idrogeno non era osservabile spettroscopicamente, ma sembrava ugualmente la componente principale più ragionevole data la sua abbondanza nell'atmosfera terrestre.

La stima di 85 millibar per la pressione era stata dedotta dall'intensità e dalla polarizzazione della luce solare riflessa, facendo uso di certe ipotesi per



La nebbia di ghiaccio imbianca di primo mattino il fondo delle depressioni della regione di Marte nota come Coprates Chasma. Nubi di ghiaccio si muovono attraverso le valli e gli altipiani circostanti. Il ripetersi giornaliero di tali nebbie suggerisce che del vapore acqueo si condensi sul terreno durante la notte formando ghiaccio, che evapora poi al sole del mattino. In seguito, il vapore rigela salendo attraverso la fredda atmosfera. La fotografia è stata ottenuta componendo tre immagini in bianco e nero riprese attraverso un filtro violetto, uno verde e uno arancio dall'*orbiter* di *Viking 1* nel luglio 1976. Questa illustrazione e quella di pagina 34 sono state preparate dall'Image Processing Laboratory del Jet Propulsion Laboratory del California Institute of Technology. Queste e altre immagini riprese dai Viking sono state fornite da Geoffrey A. Briggs del gruppo per l'elaborazione di immagini trasmesse dall'*orbiter* dei Viking.



Marte in fase crescente fu fotografato dalla sonda spaziale *Viking 2* nell'agosto 1976, da una distanza di 450 000 chilometri. La grande struttura circolare a sud, il bacino di Argyre, è coperta da un sottile strato di ghiaccio che si estende dalla calotta polare in inverno. Le configurazioni lineari più a nord fanno parte del sistema di canyon detto Valles Marineris, che si allunga per 4000 chilometri parallelamente all'equatore. I due cerchi oscuri più a ovest presso il terminatore mattutino sono gli enormi rilievi vulcanici Ascraeus Mons e Pavonis Mons. Due dense strisce di nubi si dipartono da Ascraeus Mons, estendendosi verso ovest fino a scomparire nell'ombra al terminatore dopo circa 1000 chilometri. Le ombre delle nubi indicano che le strisce si trovano a un'altezza circa pari a quella del vulcano, che svetta a più di 20 chilometri sopra il livello medio della superficie. Le nubi rivelano che sta soffiando un forte vento da est.

separare gli effetti della luce riflessa dalla superficie del pianeta da quelli della luce diffusa o riflessa dall'atmosfera. Si sa ora che la percentuale di riflessione attribuita alla componente gassosa dell'atmosfera fu molto sovrastimata, poiché non si era tenuto conto della rilevante quantità di polvere e nebbia di cui fu accertata in seguito la presenza. Sebbene la gravità sulla superficie di Marte sia solo il 38 per cento di quella presente sulla superficie terrestre, non vi sono ragioni in linea di principio che impediscano che Marte abbia un'atmosfera con una pressione superficiale elevata quasi quanto quella terrestre. L'idrogeno e l'elio sono i soli gas la cui energia cinetica dovuta a moti termici è più che sufficiente per permettere loro di sfuggire all'attrazione gravitazionale di entrambi i pianeti.

Mentre *Mariner 4* era in viaggio verso la sua meta, un'analisi di dati spettrografici migliori mostrò che la pressione superficiale su Marte doveva essere molto minore di 85 millibar. Il suo valore fu determinato quando *Mariner 4* transitò presso il pianeta nel luglio 1965. Dal punto di vista degli osservatori terrestri *Mariner 4* sparì dietro a Marte per riapparire 54 minuti dopo. Quando i segnali radio del veicolo spaziale passarono rasenti al bordo del pianeta furono rifratti dall'atmosfera. La rifrazione osservata indicava una pressione superficiale compresa tra cinque e sette millibar. I nuovi dati spettrografici avevano mostrato un quantitativo di anidride carbonica sufficiente per spiegare una pressione di tale entità, così che l'anidride carbonica era chiaramente il costituente principale dell'atmosfera di Marte. L'azoto veniva relegato al ruolo di componente minore, ammesso che ce ne fosse almeno un po'.

Azoto e gas nobili (elio, argo, neo, cripto e xeno) furono registrati da entrambi i *lander* dei Viking durante la loro discesa attraverso l'atmosfera di Marte e, in seguito, sulla superficie del pianeta. Le analisi furono eseguite con spettrometri di massa, strumenti che dapprima ionizzano un campione di gas nel vuoto e distinguono poi gli ioni risultanti a seconda della loro massa. Furono usati due spettrometri differenti: uno che operava nel vuoto approssimativo dell'alta atmosfera di Marte mentre il *lander* attraversava la regione compresa tra 200 e 120 chilometri dalla superficie, un altro, progettato soprattutto per l'analisi del suolo, in cui il vuoto era prodotto artificialmente. Anche se non furono fatte misurazioni tra 120 chilometri e la superficie, è possibile dedurre con ragionevole sicurezza la composizione dell'atmosfera nella regione non campionata.

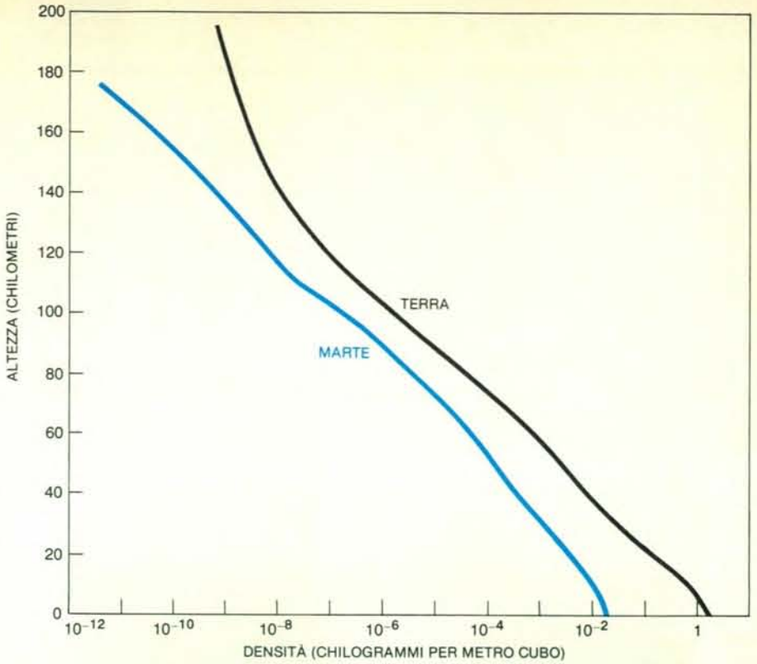
Le misure alla superficie indicano che le molecole di anidride carbonica costituiscono circa il 96 per cento del totale. L'azoto è secondo in abbondanza col 2,5 per cento, seguito dall'isotopo più comune dell'argo, argo 40, con l'1,5 per cento. In accordo con le misure precedenti fatte dalla Terra, le molecole di ossigeno costituiscono solo lo 0,1 per cento. Cripto e xeno sono pure presenti nell'atmosfera in quantità piccole ma misurabili, probabilmente insieme con tracce di neo e di elio.

Su Marte, come sulla Terra, l'atmosfera è fatta di gas liberati dall'interno del pianeta a seguito di fenomeni vulcanici e di forme meno violente di emissione. Solo una parte, però, dei gas fuorusciti è rimasta nelle due atmosfere come si osservano oggi. Sulla Terra la maggior parte delle sostanze volatili liberate si trovano ora in altri «serbatoi» o riserve. Il serbatoio che contiene la maggior parte del vapore acqueo fuoruscito è costituito, ovviamente, dagli oceani. La maggior parte dell'anidride carbonica emessa si trova ora combinata nelle rocce di carbonato di calcio, come il calcare, che si è formato sul fondo degli oceani. Altri serbatoi di carbonio, emesso originariamente come anidride carbonica, sono costituiti dai giacimenti di carbone, di petrolio e di nafta e dal carbonio disciolto negli oceani e diffuso nella biosfera. La quantità di carbonio immagazzinata in forma di anidride carbonica nell'atmosfera terrestre è trascurabile.

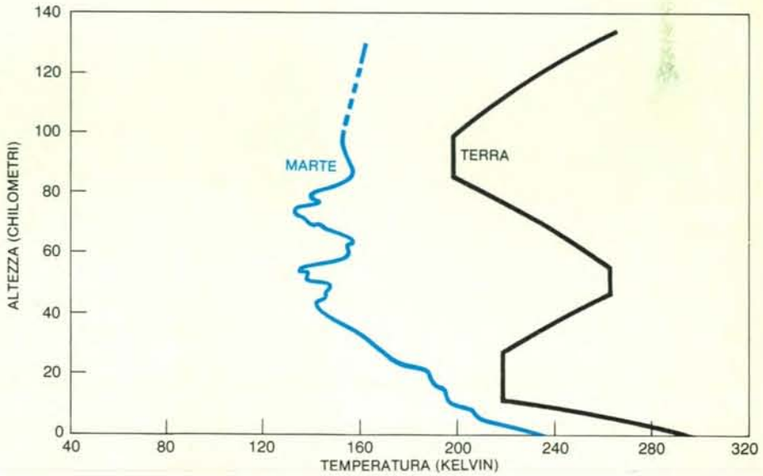
Vi possono essere su Marte anche grandi serbatoi non atmosferici di sostanze volatili fuoruscite. Una certa quantità di acqua potrebbe trovarsi nel ghiaccio polare o essere dispersa nel permafrost. È noto che dell'acqua si trova nel suolo in composti chimici. Grandi quantità di anidride carbonica possono essere assorbite nel suolo, ed è possibile che una parte si trovi congelata nella piccola calotta polare permanente dell'emisfero sud. Dato che è difficile stimare l'ammontare delle sostanze volatili accumulate in quelle riserve potenziali, occorrono ulteriori informazioni per determinare il volume totale dell'acqua e dell'anidride carbonica fuoruscite. Le abbondanze di azoto e di gas nobili presenti nell'atmosfera di Marte possono fornire tali informazioni.

Vediamo come si può procedere in proposito. Assumiamo che la Terra e Marte abbiano avuto originariamente composizioni simili, almeno per quanto riguarda carbonio, idrogeno e gas nobili. I gas nobili non solidificano e non reagiscono chimicamente, inoltre, a eccezione dell'elio, non sfuggono rapidamente dall'atmosfera. Pertanto il rapporto tra la quantità di gas nobili presenti nell'atmosfera e la quantità totale di anidride carbonica e di acqua fuoruscita durante la storia del sistema solare dovrebbe avere valori simili sia su Marte sia sulla Terra. Questo è il modello cosiddetto «ad analogia terrestre».

I gas nobili più adatti per questo confronto sono il cripto e i due isotopi rari dell'argo, argo 38 e argo 36. L'isotopo più abbondante, argo 40, è un prodotto di decadimento radioattivo del potassio 40 ed è pertanto più adatto come indicatore di presenza di potassio piuttosto che della liberazione totale di gas dall'interno del pianeta. Le misurazioni effettuate dai Viking su Marte hanno mostrato che



La densità dell'atmosfera su Marte è pari a circa un centesimo della densità dell'atmosfera terrestre a livello del suolo. Fino a un'altezza di 100 chilometri la densità dell'atmosfera terrestre diminuisce più rapidamente della densità di quella di Marte. Al di sopra di 100 chilometri la situazione si rovescia. La variazione della densità dell'atmosfera marziana al di sotto dei 100 chilometri è stata dedotta da un'analisi del comportamento del *lander* di *Viking 2* fatte da A. Seiff e da D.B. Kirk dello Ames Research Center della NASA. Al di sopra di 100 chilometri la densità atmosferica è ricavata dai dati registrati dallo spettrometro di massa posto sul *lander* di *Viking 2*. L'analisi è stata eseguita da A.O.C. Nier dell'Università del Minnesota.



La variazione della temperatura dell'atmosfera di Marte al variare della quota differisce nettamente da quella che si osserva nell'atmosfera terrestre. La sola caratteristica comune alle due atmosfere è costituita da una costante diminuzione della temperatura fino a un'altezza di 10 chilometri circa. Al di sotto di tale quota entrambe le atmosfere si riscaldano assorbendo energia dal terreno, che è riscaldato direttamente dal Sole. Il picco di temperatura nell'atmosfera terrestre a un'altezza di circa 50 chilometri è prodotto dall'assorbimento della radiazione ultravioletta da parte dell'ozono, mentre nell'atmosfera di Marte non vi è uno strato analogo di ozono che possa assorbire energia. Le irregolarità nella curva di temperatura dell'atmosfera marziana tra 50 e 100 chilometri sono dovute probabilmente al raffreddamento di strati di gas in moto ascendente e al riscaldamento di strati di gas che scendono. La temperatura di Marte a circa 130 chilometri fu ottenuta da un esperimento di ingresso del Viking analizzato da William B. Hanson dell'Università del Texas. Le altre temperature provengono da analisi condotte da Seiff e Kirk.

le abbondanze relative dei due isotopi dell'argo e i loro rapporti rispetto al cripto sono simili ai valori misurati sulla Terra, il che è coerente con il modello ad analogia terrestre. È rischioso estrapolare tale modello all'anidride carbonica ma, se lo si fa, si trova che la massa totale di anidride carbonica liberata su Marte è pari solo a dieci volte circa la quantità presente oggi nell'atmosfera. Tale quantità è sorprendentemente piccola, e ci indica, nel caso sia almeno approssimativamente corretta, che la quantità totale del gas fuoruscito su Marte è stata inferiore di molte centinaia di volte a quella che si è liberata sulla Terra.

Un modo completamente diverso per stimare il volume totale dei gas liberati attraverso la superficie di Marte consiste nell'esaminare le abbondanze isotopiche dell'azoto, dell'ossigeno e del carbonio presenti nell'atmosfera per vedere quanto si siano accresciuti gli isotopi più pesanti di quegli elementi rispetto a quelli più leggeri, che hanno maggiori probabilità di fuggire dall'atmosfera. Né l'ossigeno né l'azoto possiedono abbastanza energia cinetica da moti termici per sfuggire dall'atmosfera, ma entrambe le molecole possono sfuggire lentamente come conseguenza della ionizzazione prodotta dalla radiazione ultravioletta nell'alta atmosfera. Quando la molecola ionizzata e gli elettroni si ricombinano, l'energia liberata è più che sufficiente per dissociare le molecole negli atomi costituenti. L'energia rimasta fornisce agli atomi di ossigeno e di azoto un'energia cinetica sufficiente per permettere una lenta fuga di entrambi i gas. Lo stesso processo non ha luogo sulla Terra a causa della maggiore attrazione gravitazionale terrestre.

Le misurazioni effettuate dai Viking indicano che le abbondanze dei vari isotopi rimangono costanti fino a 120 chilometri. Al di sopra di tale quota i rapporti cominciano a cambiare, perché la turbolenza non è sufficiente per assicurare un perfetto mescolamento dei gas. Nella atmosfera estremamente rarefatta al di sopra di 120 chilometri le concentrazioni dei vari gas sono controllate dalla diffusione molecolare, un processo che fa sì che ogni gas si distribuisca secondo la sua massa molecolare: i gas più leggeri salgono come crema ai confini estremi dell'atmosfera per poi fuggire nello spazio. La zona che separa la regione inferiore ben mescolata dell'atmosfera dalla regione di diffusione è detta «turbopausa». Su Marte la turbopausa è a una altezza di circa 120 chilometri, dove la densità dell'atmosfera è pari solo a un centesimo della densità che si ha in corrispondenza della turbopausa terrestre.

La fuga verso l'esterno da un atmosfera planetaria è possibile solo per molecole che si trovino sopra la turbopausa, perché solo al di sopra di essa possono sfuggire prima di urtare altre molecole. Dato che gli isotopi più leggeri sono relativamente più abbondanti oltre la turbopausa, essi sfuggiranno più rapidamente delle loro controparti più pesanti. A lungo andare, questo processo determinerà in tutta l'atmosfera un arricchimento selettivo degli isotopi pesanti rispetto a quelli leggeri. Le misure dei Viking mostrano che il rapporto tra l'isotopo raro pesante dell'azoto, azoto 15, e l'isotopo comune, azoto 14, è maggiore di circa il 60 per cento del rapporto corrispondente sulla Terra. Il grado di arricchimento dipende dalla quantità totale di azoto fuoruscita, dal tempo trascorso dall'emissione e dall'altezza della turbopausa. Quanto più alta è la turbopausa, tanto minore è l'arricchimento dell'azoto 15 nella regione da cui è possibile la fuga. L'arricchimento dipende anche da fattori esterni quali il flusso di energia solare nell'ultravioletto, che agevola il processo di fuga fornendo l'energia necessaria per la ionizzazione. Modelli di evoluzione dell'atmosfera di Marte basati sul rapporto tra gli isotopi dell'azoto suggeriscono che l'emissione di azoto cominciò presto nella storia del pianeta e che il volume totale emesso è stato grande, forse 100 volte la quantità attualmente presente nell'atmosfera.

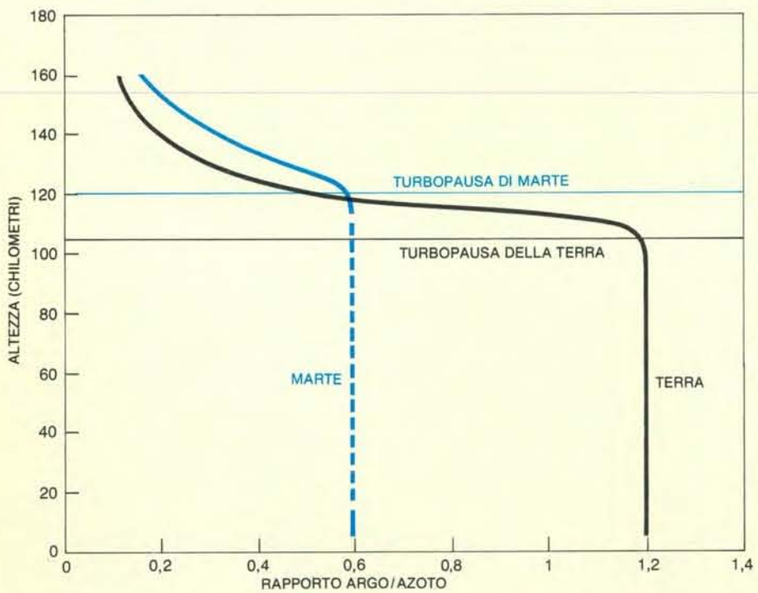
Secondo le misurazioni effettuate dai Viking, non vi è stato un arricchimento corrispondente degli isotopi più pesanti del carbonio e dell'ossigeno. Come si può spiegare questo fatto? È stato proposto che carbonio e ossigeno atmosferici siano in contatto, a differenza dell'azoto, con grandi riserve, in superficie o appena al di sotto di essa, di materiali contenenti carbonio e ossigeno, così che siano possibili scambi atomici tra l'atmosfera e la riserva. Le quantità di carbonio e ossigeno presenti nell'atmosfera potrebbero rappresentare solo la punta dell'iceberg. Pertanto, l'assenza di un arricchimento osservabile negli isotopi più pesanti del carbonio e dell'ossigeno può essere usata come indicazione che un ingente volume di gas fuoruscito ha formato una grande riserva di anidride carbonica. Sembra corretto concludere che la liberazione di gas su Marte, anche se molto meno efficiente di quella terrestre, abbia prodotto quantità di anidride carbonica e di vapore acqueo significativamente maggiori di quelle che si osservano oggi nell'atmosfera.

A differenza dell'azoto e dei gas nobili il vapore acqueo è una componente molto variabile dell'atmosfera di Marte. Misure fatte dalla Terra hanno rivelato un ciclo stagionale e un accenno di ciclo giornaliero. Gli spettrometri a bordo degli *orbiter* dei Viking hanno mostrato ora che la quantità di vapore acqueo presente varia molto con la latitudine. Sebbene sia difficile determinare spettroscopicamente la distribuzione verticale del vapore acqueo, si può calcolare rapidamente la massa totale che si estende dal suolo fino ai limiti dell'atmosfera. Convertendo al suo equivalente liquido la quantità di vapore acqueo in tale colonna si ottiene il numero di centimetri di precipitazione possibili, vale a dire l'altezza in centimetri del liquido se la precipitazione fosse possibile.

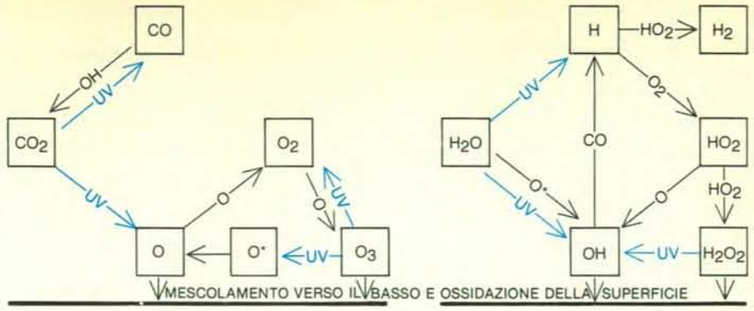
I valori misurati finora dagli *orbiter* dei Viking variano tra meno di 10^{-4} centimetri possibili di precipitazione ad alte latitudini meridionali a metà inverno e 10^{-2} centimetri precipitabili ad alte latitudini settentrionali in piena estate. L'atmosfera terrestre contiene di solito due o tre centimetri precipitabili e, pertanto, l'atmosfera di Marte sembra, a prima vista, estremamente secca. Più correttamente, però, il confronto dovrebbe essere fatto col quantitativo totale di acqua presente nell'atmosfera terrestre al di sopra del livello ove la pressione è pari alla pressione presente alla superficie di Marte: sette millibar. In base a questo criterio l'atmosfera di Marte è molto umida; perfino il valore più piccolo misurato su Marte supera di varie volte il contenuto di acqua dell'atmosfera terrestre. Una misura più significativa è quella dell'umidità relativa, ovvero del grado di saturazione dell'atmosfera. L'umidità relativa dell'atmosfera di Marte è in media così elevata che l'abbondanza di vapore acqueo sembra sia controllata soprattutto dalla saturazione. L'umidità dell'atmosfera è prossima al limite fissato dalla sua temperatura.

Occorre però chiaramente qualche altro fattore, diverso dalla saturazione, per mantenere la variazione osservata del contenuto di vapore acqueo con la latitudine. Un fattore importante è costituito dalla distribuzione dei generatori e dei distributori di vapore acqueo. In tarda estate si osserva in corrispondenza a ciascun polo una piccola calotta di ghiaccio residua, e si sa che, almeno al polo nord, la calotta è fatta di ghiaccio d'acqua. La sublimazione del ghiaccio sotto il Sole estivo sarebbe una sorgente abbondante di vapore acqueo. Non sorprende che le maggiori quantità di vapore osservate dai Viking siano apparse vicino ai bordi della calotta polare settentrionale. Stranamente, però, nessun aumento analogo fu rilevato dall'*orbiter* di *Mariner 9* attorno al residuo della calotta polare meridionale in piena estate. Se gli *orbiter* dei Viking riusciranno a raccogliere dati per un intero ciclo annuale, il mistero potrà essere chiarito.

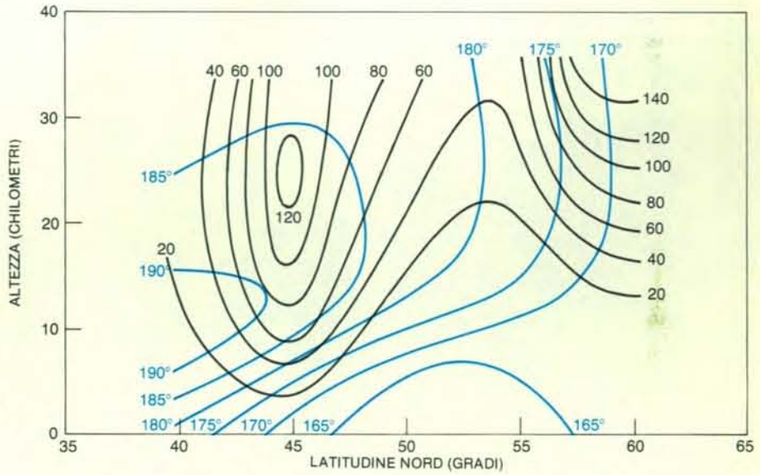
Il vapore acqueo è il componente chiave della chimica dell'atmosfera di Marte. A differenza della Terra, che è schermata da un denso strato di ozono contro la radiazione solare ultravioletta, Marte è esposto a questa radiazione energetica che può raggiungerne la superficie. A causa di ciò in tutta l'atmosfera l'anidride carbonica è dissociata in ossido di carbonio (CO) e ossigeno atomico (O), e il vapore acqueo è dissociato in idrogeno atomico (H) e radicale ossidrile (OH). L'H e l'OH sono estremamente reattivi.



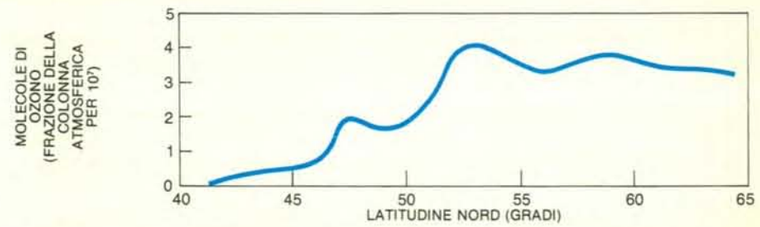
Il rapporto argo-azoto nelle atmosfere di Marte e della Terra diminuisce parallelamente al di sopra di 120 chilometri circa, ma assume valori nettamente diversi al di sotto delle rispettive turbopause dei due pianeti. La turbopausa segna la zona di transizione tra una regione inferiore dell'atmosfera, in cui i venti assicurano un buon mescolamento dei gas costituenti, e una regione superiore a bassa densità, dove i vari componenti possono essere separati dalla diffusione molecolare. Al di sopra della turbopausa il rapporto argo-azoto diminuisce rapidamente poiché l'azoto, il più leggero dei due gas, aumenta in abbondanza relativa prima della sua lenta fuga da Marte. La curva relativa a Marte al di sopra di 110 chilometri è stata ricavata da esperimenti di ingresso dei Viking. I valori presso la superficie di Marte provengono dallo spettrometro di massa sul *lander* di *Viking 2* in un esperimento di K. Biemann del Massachusetts Institute of Technology e di T.C. Owen della State University di New York a Stony Brook.



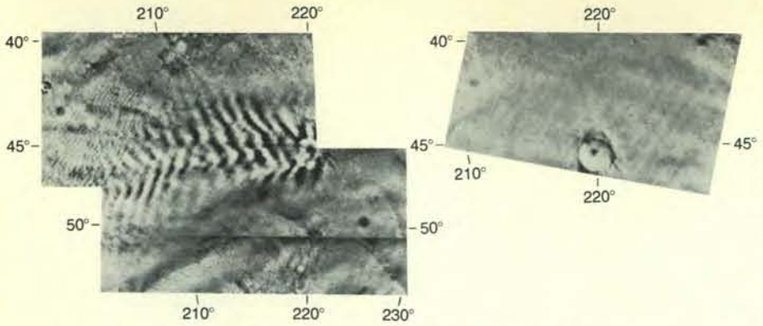
Le reazioni chimiche atmosferiche hanno inizio quando l'anidride carbonica (CO_2) e il vapore acqueo (H_2O) sono dissociati dalla radiazione ultravioletta (UV). Nel ciclo dell'ossigeno (*a sinistra*) CO_2 si scinde in ossido di carbonio (CO) e ossigeno atomico (O). Quest'ultimo si ricombina formando O_2 e ozono (O_3). Sia O che O_3 contribuiscono all'ossidazione delle rocce e del terreno. Nel ciclo dell'idrogeno (*a destra*) H_2O si scinde in idrogeno atomico (H) e nel radicale ossidrile (OH). Le reazioni tra H e O_2 danno luogo al radicale perossidrile (HO_2), che può reagire con se stesso formando perossido di idrogeno (H_2O_2). Sia H_2O_2 sia OH possono contribuire all'ossidazione della superficie. I due cicli interagiscono soprattutto attraverso le reazioni di CO con OH, di O_2 con H, e di O con HO_2. L'idrogeno atomico eccitato (O*), prodotto dalla dissociazione ultravioletta di O_3, può a sua volta spezzare H_2O. Sia H_2 che O_2 lasciano a poco a poco l'atmosfera spinti verso l'alto dal rimescolamento finché fuggono via. Gli aspetti quantitativi dei cicli qui illustrati sono stati elaborati da Michael B. McElroy della Harvard University e da Thomas M. Donahue dell'Università del Michigan e, in modo indipendente, da Truman D. Parkinson e Donald M. Hunten del Kitt Peak National Observatory.



La temperatura invernale dell'atmosfera di Marte (*in colore*), rappresentata in un piano meridiano, è stata ricavata da misure in infrarosso fatte da *Mariner 9*. Le temperature, in kelvin, possono essere usate per dedurre la distribuzione della componente est-ovest dei venti (*in nero*), espressa in metri al secondo. Questo esperimento in infrarosso di *Mariner 9* è stato condotto da R.A. Hanel, B.J. Conrath e collaboratori del Goddard Space Flight Center della NASA.



La distribuzione dell'ozono in inverno è rappresentata per lo stesso piano usato nell'illustrazione precedente. L'asse verticale mostra il numero di molecole di ozono, come frazione di tutte le molecole contenute in una colonna verticale di atmosfera. L'ozono cresce verso i poli, dove le temperature e le concentrazioni di vapore acqueo sono basse. I valori sono stati ottenuti da C.A. Barth e collaboratori all'Uuniversità del Colorado dai dati in ultravioletto di *Mariner 9*.



Nubi a onde nell'atmosfera di Marte in pieno inverno sono ben visibili nel mosaico di tre fotografie a sinistra, scattate nel 1972 da *Mariner 9*. Il lungo treno di nubi oscura quasi il grande cratere sopra al quale si sono formate. Le singole nubi della formazione a onde distano circa 20 chilometri, cioè metà del diametro del cratere. Il cratere libero da nubi appare sul bordo inferiore della fotografia singola a destra, che fu ripresa un giorno prima. Latitudine e longitudine sono indicate lungo i bordi dei mosaici. Per evitare che il rilievo del cratere e delle altre strutture superficiali apparisse invertito, le immagini sono state riprodotte col nord in basso.

Nubi a onde al riparo di un cratere di Marte furono fotografate da *Mariner 9* nel 1972. La nube a onde estesa e diffusa che appare in alto a destra è simile alle nubi che si vedono sopra il cratere di 40 chilometri nei mosaici a bassa risoluzione in alto nella pagina. Le nubi sono costituite probabilmente da ghiaccio d'acqua. Il cratere che dà origine a questa nube ha un diametro di soli sette chilometri circa. In basso a sinistra e sottovento dal cratere vi sono strisce di nubi, che, per i bordi netti e la struttura a piccola scala, sembrano non di acqua, ma di anidride carbonica solida che in inverno può congelare rapidamente nell'atmosfera sopra le regioni polari.

Per esempio, possono fare da catalizzatori per la ricombinazione di CO e O a dare di nuovo anidride carbonica, il che contribuisce a spiegare perché l'atmosfera sia fatta essenzialmente di anidride carbonica. L'idrogeno atomico e l'ossigeno molecolare si combinano attraverso una serie di reazioni producendo perossido di idrogeno (H_2O_2), un potente ossidante, che può avere un influsso importante sulla chimica del suolo.

Un piccolo quantitativo di idrogeno molecolare (H_2) è prodotto dalla dissociazione del vapore acqueo. Dato che è molto stabile, H_2 è ben mescolato sotto la turbopausa. Al di sopra della turbopausa si dissocia in idrogeno atomico, che sfugge lentamente. Se si estrapola su tutta la vita di Marte (4,6 miliardi di anni) il tasso di fuga stimato oggi, si trova che esso potrebbe spiegare la perdita di uno strato di acqua in fase liquida alto meno di un metro. Tale quantità è probabilmente molto minore del quantitativo totale di acqua fuoruscito dal pianeta.

Nonostante che le quattro specie chimiche contenenti idrogeno, H, H_2, OH e H_2O_2, siano importanti, esse sono al di sotto della soglia di sensibilità degli strumenti posti sul Viking. Le reazioni chimiche atmosferiche che ho descritto sono dedotte dall'esperienza terrestre. Comunque, gli spettrometri ultravioletti sulle sonde spaziali Mariner hanno rivelato l'idrogeno atomico che sfugge dall'alta atmosfera di Marte. Tali osservazioni sono a favore della dissociazione ipotizzata dell'idrogeno molecolare sopra la turbopausa.

La distribuzione osservata dell'ozono (O_3) nell'atmosfera di Marte fornisce una prova più consistente a favore dello schema di reazioni chimiche dedotto. Sebbene l'ozono sia molto meno abbondante che nell'atmosfera terrestre, concentrazioni di esso facilmente rivelabili si trovano in inverno sopra le regioni polari di Marte. A latitudini più basse e più calde di solito non è rivelabile. L'ozono è prodotto dalla combinazione di O e di O_2, che sono stati generati dalla dissociazione dell'anidride carbonica. L'ozono e il suo precursore, l'ossigeno atomico, sono distrutti da reazioni in cui H e OH fanno da catalizzatori. Pertanto la concentrazione di ozono sarà bassa dovunque esistano ragguardevoli quantità di H e OH, ovvero dovunque sia abbondante il materiale che li genera, cioè il vapore acqueo. Dato che il vapore acqueo è praticamente assente attorno al polo invernale, l'ozono vi si troverà abbondantemente. La limitazione che la dissociazione del vapore acqueo pone sulla concentrazione di ozono esercita un'analoga funzione nell'atmosfera terrestre. Al di sopra di 40 chilometri di altezza la concentrazione di ozono è limitata dalle medesime reazioni catalitiche con H e OH. Dunque, l'atmosfera di Marte assomiglia alla stratosfera terrestre sia per quanto riguarda la chimica sia per quanto riguarda la pressione.

Dato che l'atmosfera di Marte è prossima alla saturazione, ci si aspetterebbe che le nubi di ghiaccio siano abbondanti, come in effetti appaiono almeno durante la stagione estiva settentrionale che le sonde Viking hanno esaminato finora. Sempre in accordo con le aspettative, le nubi di ghiaccio d'acqua sono sottili, con bordi diffusi. Non c'è nulla che assomigli alle nubi cumuliformi dense, a limiti ben definiti, che conosciamo sulla Terra. Le nubi marziane sono di quattro tipi generali: nubi convettive, nubi a onde, nubi orografiche e nebbie.

Le nubi convettive si producono quando il gas presso la superficie è riscaldato durante il giorno, sale e si raffredda a causa dell'espansione. Quando la temperatura scende fino al punto di saturazione, le nubi si formano in fiocchi distinti di dimensioni e spaziatura quasi uniformi. Queste strutture di nubi sono frequenti verso mezzogiorno sulle regioni montuose equatoriali. La distanza fra i fiocchi e l'altezza delle nubi stimate dalla posizione delle loro ombre indicano che un'attiva convezione ha luogo in un livello alto da cinque a otto chilometri, un po' più alto dello strato di riscaldamento diurno sopra le aree continentali terrestri. L'importante rimescolamento giornaliero prodotto dalla convezione assicura che tutti i componenti atmosferici presenti in tale livello possano interagire efficacemente con la superficie di Marte.

Le immagini migliori di nubi a onde furono riprese da *Mariner 9,* che esaminò l'emisfero settentrionale in inverno. Le nubi a onde si formano quando un vento forte soffia costantemente sopra un ostacolo, come una cresta elevata. Si formano allora, al riparo dell'ostacolo, treni di onde atmosferiche, simili a onde nell'acqua. Se le condizioni di umidità e di temperatura sono adatte, le creste delle onde sono accompagnate da nubi che sembrano stazionarie.

Le nubi a onde su Marte indicano che la meteorologia delle regioni a media latitudine presenta un'evidente somiglianza con quella delle zone corrispondenti sulla Terra. I venti dominanti soffiano da ovest a ogni quota fino a un'altezza notevole, e i venti più forti, come le correnti a getto, si trovano vari chilometri sopra la superficie. Combinando dati di temperatura con le fotografie delle nubi a onde si può ricostruire un'immagine tridimensionale della distribuzione dei venti in certi giorni. Di solito il vento soffia da ovest con velocità che va dai 10-20 metri al secondo presso la superficie fino a più di 100 metri al secondo (360 chilometri all'ora) alla quota delle correnti a getto, sopra i 10 chilometri. Queste velocità sono circa doppie di quelle che si osservano ai livelli corrispondenti sulla Terra.

Le nubi orografiche si formano quando il gas atmosferico è costretto a salire lentamente e costantemente, come quando si muove salendo le pendici di un grande rilievo montuoso. Tali nubi sono comuni sopra le regioni vulcaniche più elevate di Marte. Nella stagione estiva settentrionale, quando l'atmosfera è umida, grandi aree sono coperte da nubi sottili piuttosto uniformi, osservabili dalla Terra. Per giorni di seguito il loro schema di crescita e decadimento si ripete con notevole precisione, indicando la regolarità delle condizioni atmosferiche giorno dopo giorno.

Una delle scoperte atmosferiche più sorprendenti fatta dagli *orbiter* dei Viking fu l'osservazione delle nebbie che si generano al suolo di primo mattino in varie regioni a bassa quota. Chiaramente, le nebbie si originano quando il gelo del terreno è trasformato in vapore dal Sole mattutino. Il vapore si condensa nuovamente man mano che sale attraverso la fredda atmosfera del primo mattino. Dato che le nebbie si osservano giorno dopo giorno nello stesso luogo, sembra probabile che il vapore acqueo atmosferico geli nuovamente sulla superficie ogni notte.

L'acqua non è l'unica sostanza capace di generare nubi su Marte. In inverno nelle regioni polari e ad alte latitudini la temperatura può scendere quanto basta perché l'anidride carbonica condensi producendo nubi di ghiaccio secco. Dato che la condensazione e la sublimazione dell'anidride carbonica possono essere rapide, le nubi di ghiaccio secco, a differenza di quelle di ghiaccio d'acqua, possono avere bordi netti. Si pensa che varie nubi brillanti, a bordi ben definiti, fotografate dalle sonde Mariner e Viking, siano fatte di ghiaccio secco. È pure possibile che in inverno vi siano tempeste di neve di ghiaccio secco che contribuiscono a generare le calotte polari stagionali di ghiaccio secco. Gran parte del ghiaccio secco delle calotte si deposita, però, probabilmente quando l'anidride carbonica gassosa viene a diretto contatto con il freddo suolo di Marte e si condensa.

Circa il 20 per cento dell'anidride carbonica nell'atmosfera di Marte è soggetto ogni stagione a un ciclo di scambio tra l'atmosfera e le calotte polari invernali, causando una corrispondente variazione nella pressione atmosferica. La variazione di pressione, che è avvertita in tutta l'atmosfera, e non solo nelle regioni delle calotte polari, è stata misurata dai barometri sui *lander* di entrambi i Viking. L'ampiezza della variazione di pressione è in buon accordo con i modelli teorici pre-Viking sullo scambio di anidride carbonica tra le calotte polari e l'atmosfera. Le misure indicano che qualsiasi altra grande riserva di anidride carbonica, per esempio l'anidride carbonica dispersa nel suolo di Marte, non è soggetta a scambi stagionali con l'atmosfera. È però possibile che esista una grande riserva di anidride carbonica assorbita capace di scambi con l'atmosfera su scale temporali molto più ampie.

I venti di Marte sono da tempo oggetto di indagine. In anni recenti sono stati fatti seri sforzi per sviluppare modelli teorici di essi con tecniche simili a quelle usate per le previsioni del clima e delle condizioni atmosferiche sulla Terra. Tali tecniche dovrebbero essere applicabili dato che entrambi i pianeti mostrano una successione simile di stagioni (poiché i loro assi polari hanno un'inclinazione quasi identica) e circa la stessa velocità di rotazione giornaliera (24 ore e 37 minuti per Marte). La velocità di rotazione intorno al proprio asse è importante perché determina l'ampiezza della deviazione delle masse atmosferiche in movimento. Tale deflessione, o effetto di Coriolis, è la causa della rotazione antioraria dei venti attorno alle aree di bassa pressione nell'emisfero nord e della rotazione oraria attorno alle aree di bassa pressione nell'emisfero sud.

Il quadro che emerge dagli studi teorici è un modello planetario di venti con due regimi principali. Uno è un regime invernale a medie latitudini, simile all'analogo regime terrestre, con venti prevalenti da ponente, perturbazioni in movimento e correnti a getto ad alta quota, o tempeste. L'altro è un regime equatoriale estivo privo di perturbazioni in movimento, privo in effetti di qualsiasi variazione da un giorno all'altro, tranne il cambiamento graduale associato al lento trascorrere delle stagioni. Nel secondo regime l'azione principale sul flusso dei venti è dovuta alla variazione giornaliera del riscaldamento solare e alla sua interazione con la topografia locale. Pertanto il gas riscaldato si muoverebbe verso l'alto lungo le pendenze locali durante il giorno e scenderebbe di nuovo durante la notte. Dato che su Marte il terreno presenta pendenze quasi ovunque, i venti che si generano a causa della pendenza del terreno, deflessi dalla rotazione planetaria, dovrebbero essere diffusi un po' dappertutto.

Inoltre il ciclo di riscaldamento giornaliero dovrebbe produrre maree atmosferiche, ovvero oscillazioni su scala planetaria del modello dei venti. Dato che Marte è un pianeta desertico, le maree dovrebbero essere più pronunciate di quelle terrestri. Presso la superficie i venti di marea sono deboli: sulla Terra raggiungono velocità al più di un metro al secondo; su Marte durante la stagione estiva dell'emisfero settentrionale dovrebbero arrivare forse a cinque o 10 volte tale velocità. Le velocità crescono però rapidamente con la quota. Sulla Terra raggiungono i 50 metri al secondo circa ad altezze comprese tra 80 e 100 chilometri. Sebbene non siano state misurate le velocità corrispondenti su Marte, si può dedurre dall'altezza della turbopausa che sono più elevate. I venti di marea sono agenti efficaci di rimescolamento nell'alta atmosfera, così che probabilmente controllano l'altezza della turbopausa su entrambi i pianeti. Dato che la turbopausa marziana è a un livello di densità minore di quello a cui si trova la turbopausa terrestre, come conseguenza di un rimescolamento più efficace da parte dei venti, sembra probabile che le maree atmosferiche presenti su Marte siano più intense di quelle che si hanno sulla Terra.

Le osservazioni dei Viking forniscono altre prove a sostegno di questa ipotesi? Forse la prova più convincente dell'intensità delle maree atmosferiche su Marte proviene dalle registrazioni barometriche nei due luoghi di atterraggio dei Viking. Gli strumenti rivelano ampie variazioni giornaliere nella pressione atmosferica, una chiara indicazione di intense maree atmosferiche. Altre indicazioni favorevoli sono state fornite dalle misurazioni della variazione della temperatura con la quota eseguite dalle sonde in avvicinamento. A quote più elevate, dove i venti di marea dovrebbero essere molto intensi, si dovrebbero osservare ampie escursioni termiche causate dall'alternarsi di livelli di gas atmosferico riscaldato per compressione e di gas raffreddato per espansione. La teoria prevede che le escursioni si completino in 20 chilometri e che la loro ampiezza cresca al crescere della quota.

I profili di temperatura registrati durante l'avvicinamento dei *lander* dei Viking mostrano escursioni termiche proprio con tali caratteristiche. Sembra dunque che su Marte, come sulla Terra, l'altezza della turbopausa sia determinata dall'intensità delle maree atmosferiche. La turbopausa agisce a sua volta come una specie di valvola che regola la fuga dei gas leggeri. Dato che la turbopausa marziana è a una quota relativamente elevata, il volume dei gas leggeri disponibile per la fuga è corrispondentemente piccolo, ed è in parte per questa ragione che il tasso odierno di fuga dell'idrogeno su Marte è così ridotto.

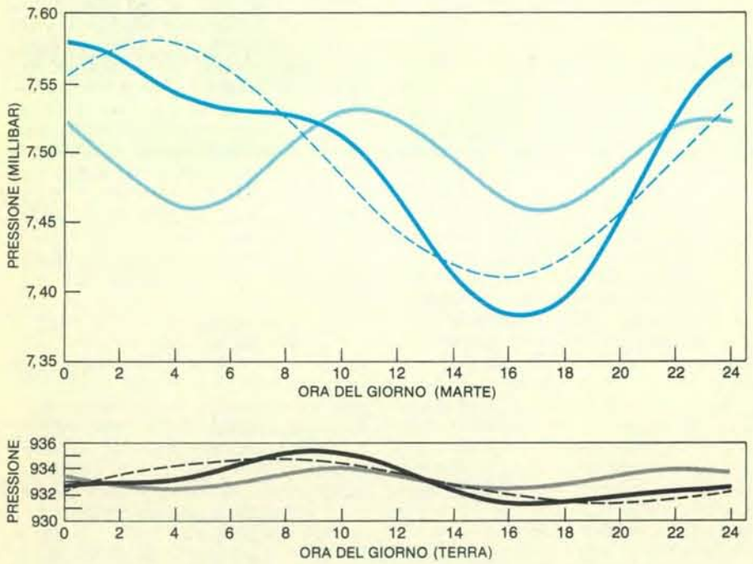
Le grandi tempeste di polvere su Marte furono osservate molto prima dell'epoca delle esplorazioni spaziali. A causa della bassa densità dell'atmosfera di Marte, i venti necessari per sollevare tempeste devono essere molto più forti di quelli che occorrerebbero sulla Terra. Un vento con una velocità di soli sei o sette metri al secondo (da 22 a 25 chilometri all'ora) è sufficiente per sollevare granelli di sabbia in molti deserti terrestri. Studi di laboratorio mostrano che su Marte sono necessari venti con velocità almeno da 30 a 60 metri al secondo (da 108 a 216 chilometri all'ora). Dato che non c'è pioggia su Marte che ripulisca l'aria, piccole particelle di polvere possono rimanere sospese per settimane o anche per mesi.

Venti abbastanza intensi da provocare tempeste di polvere si sviluppano in inverno nella fascia priva di neve presso il bordo della calotta polare marziana. Numerose tempeste di sabbia furono fotografate in quella regione sia da *Mariner 9* sia dai Viking. Molto più impressionanti sono le tempeste di polvere su scala planetaria, una delle quali avvolse completamente il pianeta quando *Mariner 9* arrivò e si mise in orbita nel novembre 1971. Prima delle missioni Viking si credeva che le tempeste a carattere planetario si ripetessero regolarmente, una volta per ogni anno marziano, più o meno quando il pianeta giunge alla minima distanza dal Sole. Molto recentemente, però, le sonde spaziali Viking hanno osservato tempeste di polvere su scala quasi planetaria in grande anticipo sulla stagione marziana prevista. Evidentemente, le tempeste a larga scala sono più comuni di quanto si credesse. Le tempeste planetarie cominciano con una o più tempeste a piccola scala nei tropici meridionali, si sviluppano poi rapidamente così da coprire la maggior parte del pianeta in pochi giorni.

Che cosa può generare i forti venti necessari per queste enormi tempeste? Il fatto che l'orbita di Marte sia considerevolmente più ellittica di quella terrestre deve avere un peso significativo. La quantità di energia solare incidente su Marte quando il pianeta è alla minima distanza dal Sole è più grande del 40 per cento circa di quella incidente quando il pianeta è alla massima distanza. La differenza analoga per la Terra è pari al 3 per cento solamente. Questo fatto non può, comunque, giustificare da sé le velocità dei venti richieste.

Sono state proposte varie spiegazioni, ma discuterò solo quella che ritengo più plausibile. L'intensità dei venti di marea su scala planetaria è molto sensibile al riscaldamento dell'atmosfera. Tale riscaldamento dipende a sua volta non solo dalla quantità di energia solare che raggiunge la superficie di Marte, ma anche dal grado di polverosità dell'atmosfera marziana: le particelle di polvere assorbono la luce solare e possono riscaldare il gas circostante direttamente ed efficacemente. Se l'atmosfera è sufficientemente carica di polvere, i venti di marea su scala globale possono crescere di intensità fino a raggiungere una forza quasi sufficiente per sollevare polvere essi stessi. Presumibilmente questo processo autorigenerativo ha inizio su aree localizzate, dove i venti di marea e quelli prodotti dalla topografia locale si combinano generando velocità abbastanza elevate da immettere una quantità critica di polvere nell'atmosfera locale. Man mano che la polvere è sparsa dai venti, cresce il riscaldamento dell'atmosfera su una grande area, producendo venti di marea di intensità sufficiente per sollevare polvere su aree ancora più vaste.

Sembra chiaro che la polvere sospesa influenza il comportamento dell'atmosfera su Marte più di qualsiasi agente noto sulla Terra. Perfino durante la stagione dei recenti atterraggi dei Viking, una stagione in cui il carico di polvere dell'atmosfera era piccolo, il cielo aveva una tonalità giallo-rosata nelle immagini riprese dai *lander* - prova evidente della presenza nell'atmosfera di una quantità significativa di polvere capace di diffondere la luce. Inoltre, sembra necessario ricorrere al riscaldamento solare di una atmosfera carica di polvere per spiegare le grandi escursioni giornaliere della pressione superficiale registrate nei luoghi di atterraggio. Sembra dunque che sia presente un efficace sistema di *feedback* positivo i cui componenti sono costituiti dai venti, dalla polvere e dal riscaldamento atmosferico.



Le variazioni giornaliere della pressione atmosferica, relativamente alla pressione totale, nel luogo di atterraggio di *Viking 1* sono circa cinque volte più grandi cha a China Lake sulla Terra. La curva continua in colore in alto mostra che nel corso di un giorno marziano la pressione atmosferica varia di 0,2 millibar, cioè del 2,5 per cento circa della pressione totale media che è di 7,5 millibar. La curva può essere decomposta in un'armonica giornaliera (*curva tratteggiata*) e in un'armonica semigiornaliera (*curva in colore leggero*). La curva nera in basso mostra la variazione giornaliera della pressione a China Lake, a sua volta decomposta in modo analogo in due armoniche. Per pareggiare la variazione giornaliera su Marte, un barometro sulla Terra dovrebbe presentare una variazione di 23 millibar nel corso di una giornata. L'escursione effettiva a China Lake è solo di 4,5 millibar. Le ampie fluttuazioni giornaliere su Marte indicano che le maree atmosferiche e i sistemi di venti sono molto più intensi che sulla Terra.

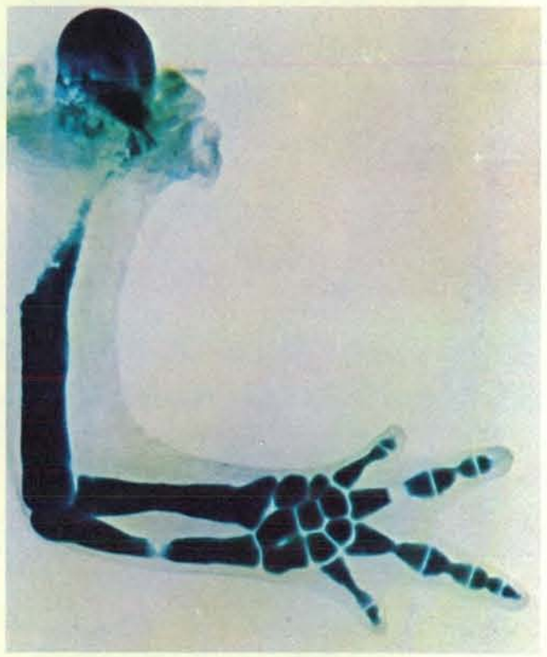
Una tempesta di polvere di grandi dimensioni su Marte fu fotografata dall'*orbiter* di *Viking 2* nel febbraio di quest'anno da un punto a 33 000 chilometri sopra la superficie del pianeta e a 51 gradi di latitudine sud con la macchina da ripresa puntata verso ovest, nord-ovest. Il centro dell'immagine è a 42,5 gradi di latitudine sud e 108 gradi di longitudine ovest, così che la tempesta ha luogo nella regione presso Claritas Fossae e Thaumasia Fossae. In pochi giorni la tempesta si estese su gran parte dell'emisfero sud e durò per varie settimane. Quando la tempesta incominciò, si era all'inizio della primavera marziana. Prima di queste osservazioni delle sonde Viking, si riteneva che le tempeste di polvere su scala quasi planetaria fossero limitate alla stagione attorno al solstizio estivo meridionale.

Sebbene rimangano delle incertezze, il quadro attuale dei processi operanti nell'atmosfera di Marte è molto dettagliato nonostante che la storia dell'esplorazione ravvicinata del pianeta sia così breve. Le immagini e le misure dei Mariner e dei Viking ci hanno rivelato molte evidenti analogie tra l'atmosfera di Marte e quella terrestre. La chimica, i processi dinamici e l'evoluzione atmosferica sono strettamente connessi su entrambi i pianeti. Gli sforzi per comprendere questi processi interagenti su un altro pianeta contribuiscono a sviluppare in profondità e in ampiezza gli studi del nostro stesso pianeta.

Possiamo dunque dire qualcosa sul clima di Marte nel passato? In particolare, i grandi canali naturali e le altre manifestazioni di erosione da acqua possono indicare che Marte ebbe in passato un lungo periodo di clima caldo e umido? Io credo che le informazioni raccolte indichino che ciò è improbabile. La quantità totale di gas liberata su Marte sembra sia stata molto minore di quella liberata sulla Terra. Anche se la massa totale liberata fosse 100 volte la massa atmosferica odierna o di più ancora, la disponibilità di vapore acqueo nell'atmosfera non poteva essere molto maggiore di quella attuale. La limitazione sulla quantità di vapore acqueo nell'atmosfera di Marte è posta dalla bassa temperatura. Anche un'atmosfera con una densità pari quasi a quella dell'atmosfera terrestre, ma composta soprattutto di anidride carbonica e di azoto, non avrebbe potuto essere molto più calda di quella odierna, e pertanto non avrebbe potuto contenere una quantità molto maggiore di vapore acqueo.

Si può immaginare che il Sole stesso fosse più caldo quando l'acqua generò per erosione le strutture che osserviamo oggi su Marte, ma più probabilmente tali strutture furono prodotte da cataclismi come impatti di meteoriti ed esplosioni vulcaniche. La crosta di Marte ha contenuto probabilmente per molto tempo grandi quantità d'acqua, o gelata o combinata chimicamente, così che qualsiasi riscaldamento improvviso avrebbe potuto provocare allagamenti locali. Un simile cataclisma avrebbe potuto produrre anche tempeste locali brevi e intense, in cui l'improvvisa vaporizzazione dell'acqua fosse seguita rapidamente da piogge torrenziali. È possibile che analoghi allagamenti a carattere di cataclisma si siano verificati anche nei primissimi stadi della storia della Terra, prima che si formassero gli oceani. È probabile che si debba attendere la prossima fase dell'esplorazione di Marte per avere risposte più precise al problema del clima esistente su Marte nel passato e a quello dell'origine dei nuovi famosi canali.

Documentazione fotografica della rigenerazione d'una zampa di blatta. Una zampa normale, come quella a sinistra, è stata amputata tra femore e trocantere: alla muta si è rigenerata una nuova zampa (*a destra*) identica alla normale, ma con un segmento tarsale in meno.



La rigenerazione dell'arto anteriore del tritone è stata evidenziata con un colorante che impregna solo la cartilagine. La zampa a sinistra è normale, con la cartilagine confinata alle zone articolari. Quando la zampa del tritone viene amputata a livello dell'omero, rigenera la parte mancante in circa sei settimane (*a destra*). Dapprima lo scheletro rigenerato è costituito unicamente da cartilagine (*in colore blu*).

La rigenerazione biologica

Le ricerche compiute su animali in grado di rigenerare nuovi arti o di sviluppare arti soprannumerari possono servire a chiarire i principi basilari che regolano l'organizzazione e la crescita delle strutture

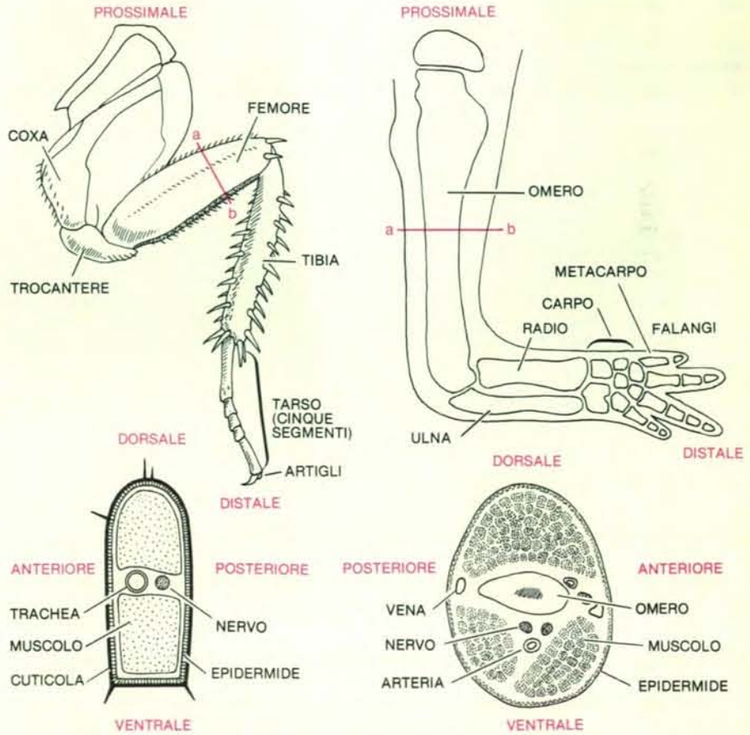
di Peter J. Bryant, Susan V. Bryant e Vernon French

Uno dei fenomeni più misteriosi della vita animale è il fatto che in ogni atto generativo l'unione di un uovo e uno spermatozoo dà origine alla formazione di un embrione provvisto di un'organizzazione spaziale prevedibile: in altre parole, le cellule embrionali si differenziano in tipi diversi di cellule organizzate secondo modelli riconoscibili e riproducibili per formare tessuti e organi specifici. Come si formino questi modelli spaziali è uno degli enigmi principali della biologia dello sviluppo: tuttavia gli embriologi stanno cominciando a comprendere che il processo deve implicare meccanismi per mezzo dei quali le cellule riescono ad accertare la loro posizione spaziale all'interno di popolazioni di cellule in via di sviluppo. L'idea che le cellule possano ricavare informazioni sulla propria posizione spaziale e reagiscano a essa differenziandosi in modo appropriato, è stata formulata da Lewis Wolpert del Middlesex Hospital Medical School di Londra.

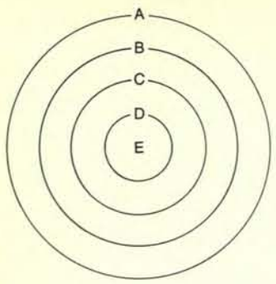
Un indizio circa il modo con cui l'organizzazione spaziale si sviluppa, ci viene fornito dalla scoperta che gli animali in via di sviluppo sono suddivisi in «territori» o «campi» in grado di svilupparsi autonomamente, almeno fino a un certo grado; questo concetto fu sostenuto in particolare da Paul A. Weiss. Un territorio embrionale può essere definito come una regione entro la quale il destino di aree diverse può essere mutato intervenendo chirurgicamente: in altri termini, i territori sono regioni in cui può avvenire una regolazione embrionale. Per esempio, come ha dimostrato Hans A. E. Driesch alla fine del XIX secolo per i ricci di mare, si possono asportare parti di embrioni agli stadi iniziali di vari animali e le parti rimanenti sono in grado di regolarsi embriologicamente in modo da formare un organismo normale. Da ciò deriva la possibilità di classificare l'embrione iniziale come un singolo territorio embrionale primario.

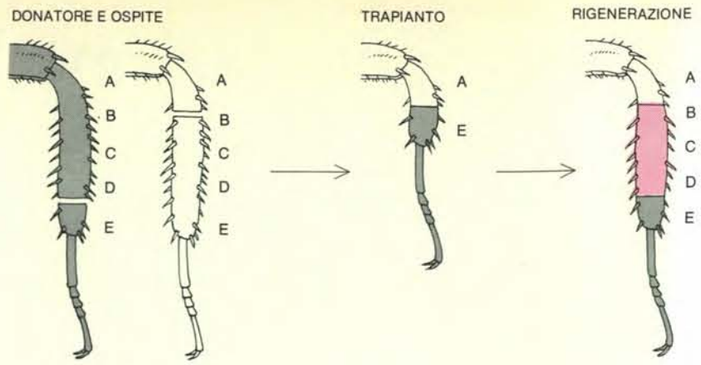
Con il procedere dello sviluppo, l'intervento chirurgico ha effetti molto più localizzati: per esempio, all'inizio di questo secolo Ross G. Harrison dimostrò che cosa accade negli embrioni degli anfibi quando tutta la regione destinata a formare l'arto anteriore e la spalla viene asportata da un embrione e trapiantata su un secondo embrione. Al termine dello sviluppo embrionale l'embrione donatore mancherà di una delle sue quattro zampe, mentre l'embrione ospite sarà dotato di cinque zampe. Se viene rimossa però solo una parte della regione embrionale, il tessuto rimanente compensa il deficit e l'arto si sviluppa.

Studi analoghi dimostrano che la zona che presiede allo sviluppo degli arti si rende indipendente dal resto dell'embrione: queste regioni indipendenti sono conosciute col nome di «territori o campi embrionali secondari». In queste zone le

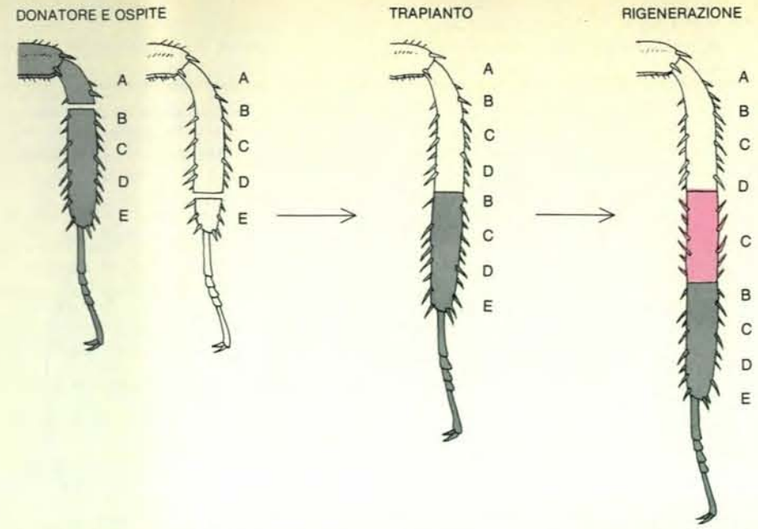


Per facilitare la comprensione delle zampe fotografate nella pagina a fronte, in questa illustrazione è stata schematizzata l'anatomia della zampa di una blatta e quella di un tritone. Entrambi gli arti sono rappresentati come se fossero visti dall'alto e poi in sezione trasversale lungo l'asse indicato (*a-b in colore*). Il tarso, ossia il segmento di zampa di blatta più distale, è costituito normalmente da cinque pezzi, come appare nella figura. L'arto anteriore del tritone possiede, invece, la sequenza ossea tipica degli anfibi e degli altri vertebrati tetrapodi, dall'omero alle falangi.





L'organizzazione d'un segmento di zampa nella blatta, come si è potuto osservare nei due esperimenti illustrati, fa pensare a una disposizione ordinata. I cerchi concentrici permettono di schematizzare tale disposizione lungo un asse, da *A*, il livello più prossimale (più vicino al centro del corpo), a *E*, il più distale. Quando, nel primo esperimento, una porzione di livello *E* della tibia viene asportata dal donatore e trapiantata sulla porzione *A* del moncherino di tibia dell'ospite, nel punto di giunzione del trapianto viene prodotta una crescita locale, con rigenerazione intercalare dei livelli d'organizzazione mancanti (in questo esempio *B*, *C* e *D*). Quando, nel secondo esperimento, una porzione di livello prossimale di tibia di un donatore viene innestata in una porzione di livello distale della tibia di un ospite, nella zona di contatto, si ha tessuto d'intercalazione, ma si forma un livello d'organizzazione capovolto (nell'esempio *C* tra *D* e *B*).



conseguenze di una manipolazione chirurgica rimangono limitate alle zone stesse. Weiss scoprì che, con il procedere dello sviluppo, i territori primari vengono sostituiti da sempre più numerosi territori o campi secondari. Nei vertebrati tali campi secondari comprendono regioni destinate alla formazione delle zampe, del cuore, degli orecchi e degli occhi. Negli insetti tra i campi secondari vi sono regioni destinate a diventare ali e zampe. Le proprietà caratteristiche dei campi secondari in alcuni animali persistono anche dopo lo sviluppo embrionale, rendendo quindi possibile la rigenerazione di parti del corpo.

In un trattato sulla rigenerazione, pubblicato nel 1901, Thomas Hunt Morgan riconobbe due diverse reazioni fondamentali degli animali all'amputazione di parti del corpo: nel primo tipo di reazione la rimanente porzione di territorio embrionale viene rimodellata e forma un organo completo ma in miniatura. Questo processo, definito morfallassi, non è legato necessariamente a una proliferazione cellulare. Nell'altro tipo di reazione, definito epimorfosi, le cellule proliferano e durante la crescita del tessuto si aggiungono nuovi elementi modellatori. Si può perciò stabilire come regola generale che i territori primari, come gli embrioni iniziali, di solito si regolano per morfallassi, mentre i territori secondari si rigenerano di solito mediante il processo di epimorfosi.

La struttura definitiva dell'embrione si sviluppa quindi da un territorio primario iniziale, tipico dell'uovo e dei primi stati embrionali attraverso lo stabilirsi di territori secondari (campi blastematici), costituiti da gruppi di cellule che occupano una determinata posizione nel territorio primario. Qui esamineremo alcuni processi mediante i quali i territori secondari possono cominciare a costituirsi nei primi stadi embrionali; tuttavia la maggior parte di ciò che conosciamo circa questo processo ci viene da studi compiuti su stadi più avanzati dello sviluppo, operando su territori secondari stabilizzati.

Prendiamo in considerazione ora in due animali, la blatta e il tritone, la notevole capacità di rigenerazione degli arti perduti. Per quanto riguarda la zampa dell'insetto, ci occuperemo del modello di sviluppo della cuticola, ossia del rivestimento chitinoso esterno prodotto da un singolo strato di cellule epidermiche. Nella zampa dell'anfibio fisseremo l'attenzione sulla struttura delle ossa, derivanti dalle cellule mesodermiche della zampa. Queste appendici sono completamente diverse ma mostrano una notevole somiglianza di risposta nei vari esperimenti di amputazione e di trapianto. Analizzando queste risposte abbiamo elaborato un modello abbastanza semplice per spiegare la regolazione di tali campi epimorfici. Dimostreremo come questo modello sia valido anche per spiegare la regolazione nella *Drosophila*.

Rigenerazione delle zampe

Quando si amputa in un qualsiasi punto la zampa di una blatta o di un tritone, per prima cosa la ferita rimargina. Le cellule al termine del moncherino perdono il loro caratteristico aspetto differenziato e si accumulano per formare la massa di tessuto noto con il nome di blastema di rigenerazione. Il blastema cresce e forma tutte le strutture che si trovano normalmente in posizione distale rispetto al punto di amputazione (il termine «distale» si riferisce a un punto più lontano dal centro del corpo, «prossimale» a un punto più vicino). Il tratto rigenerato e il moncherino formano insieme un arto completo e funzionale.

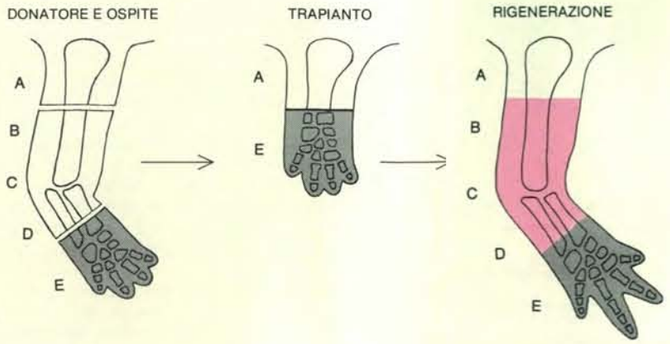
Molti ricercatori hanno dimostrato che tali strutture possono essere rigenerate anche a partire dalla superficie tagliata dell'arto amputato. Per esempio, se una porzione di zampa viene ritrapiantata in un medesimo animale, capovolgendo lo asse prossimale-distale e quindi permettendo alle parti distali, che sarebbero eliminate, di sopravvivere all'amputazione, la rigenerazione all'estremità tagliata forma tutte le strutture normalmente distali rispetto a quel punto: il risultato è una duplicazione speculare delle parti distali che sono state trapiantate con orientamento capovolto. Con una diversa operazione di trapianto, Desiré Bullière dell'Università di scienze e medicina di Grenoble e uno degli autori di questo articolo (French) hanno dimostrato che, se una giunzione di trapianto in una zampa di blatta non si cicatrizza correttamente, sia il componente del donatore, sia quello dell'ospite possono rigenerare strutture distali: ne risulta una zampa che possiede tre porzioni di parti distali.

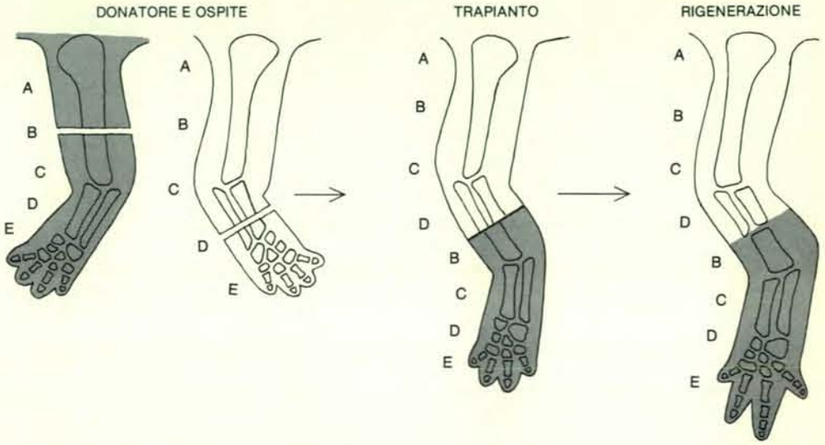
Tra il 1920 e il 1930 Hans Przibram dell'Università di Vienna formulò l'ipotesi che le bizzarre zampe multiple, che talvolta crescono spontaneamente negli insetti e negli anfibi, siano il risultato di un processo simile. Secondo il punto di vista di questo studioso una zampa parzialmente spezzata può rigenerare copie soprannumerarie della porzione di arto che si trovano oltre il punto di rottura, su entrambe le superfici rimaste esposte. Poiché la porzione della zampa originaria che si trova oltre il punto di rottura non viene eliminata, le due parti distali soprannumerarie sono costrette a crescere lateralmente.

Interazione prossimale-distale

La scoperta delle leggi che regolano il comportamento delle cellule provenienti da differenti regioni di un territorio, quando vengono costrette a interagire, ci hanno permesso di approfondire le nostre cognizioni sulla rigenerazione e sull'organizzazione spaziale delle zampe di blatta e tritoni. Per esempio, Horst Bohn, dell'Università di Monaco di Baviera, e Bullière hanno dimostrato che quando diversi livelli prossimali-distali di un segmento di zampa di blatta, come la tibia, vengono trapiantati insieme, si verifica una crescita localizzata nel punto di giunzione del trapianto. Il nuovo tessuto forma le strutture che si trovano in una zampa normale fra i livelli prossimali-distali del moncherino e il trapianto. Questo tipo di proliferazione di tessuto viene chiamato rigenerazione intercalare.

Il trapianto fra un livello prossimale di una zampa di blatta ospite e un livello distale di zampa di donatore dà origine a una rigenerazione intercalare normalmente orientata e forma un normale segmento di zampa. Che cosa si può dire, invece, a proposito di una combinazione rovesciata? Bohn e Bullière hanno trapiantato zampe amputate al donatore a livello prossimale su moncherini di un ospite amputati a livello distale. Ne risulta la medesima crescita localizzata e si forma una rigenerazione intercalare. Le nuove strutture mostrano tuttavia un orientamento rovesciato prossimale-distale. Benché il segmento risultante non mostri discontinuità, esso non si può certo definire normale. Entrambi i ricercatori hanno pure dimostrato che l'organizzazione prossimale-distale si ripete in ciascun segmento di zampa di blatta. Per esempio, trapiantando la parte intermedia di una tibia su quella intermedia di un femore non si produce la rigenerazione intercalare della parte mancante di tibia e di femore. Il trapianto della parte prossimale di una tibia sulla parte distale di un femore, d'altra parte, provoca la formazione di una rigenerazione intercalare nel punto d'innesto riproduce i livelli mancanti. Quando invece, come a destra, una zampa amputata a livello prossimale viene impiantata su un moncherino tagliato a livello distale, non viene interposto il livello tra *B* e *D* come avviene nella tibia di blatta.

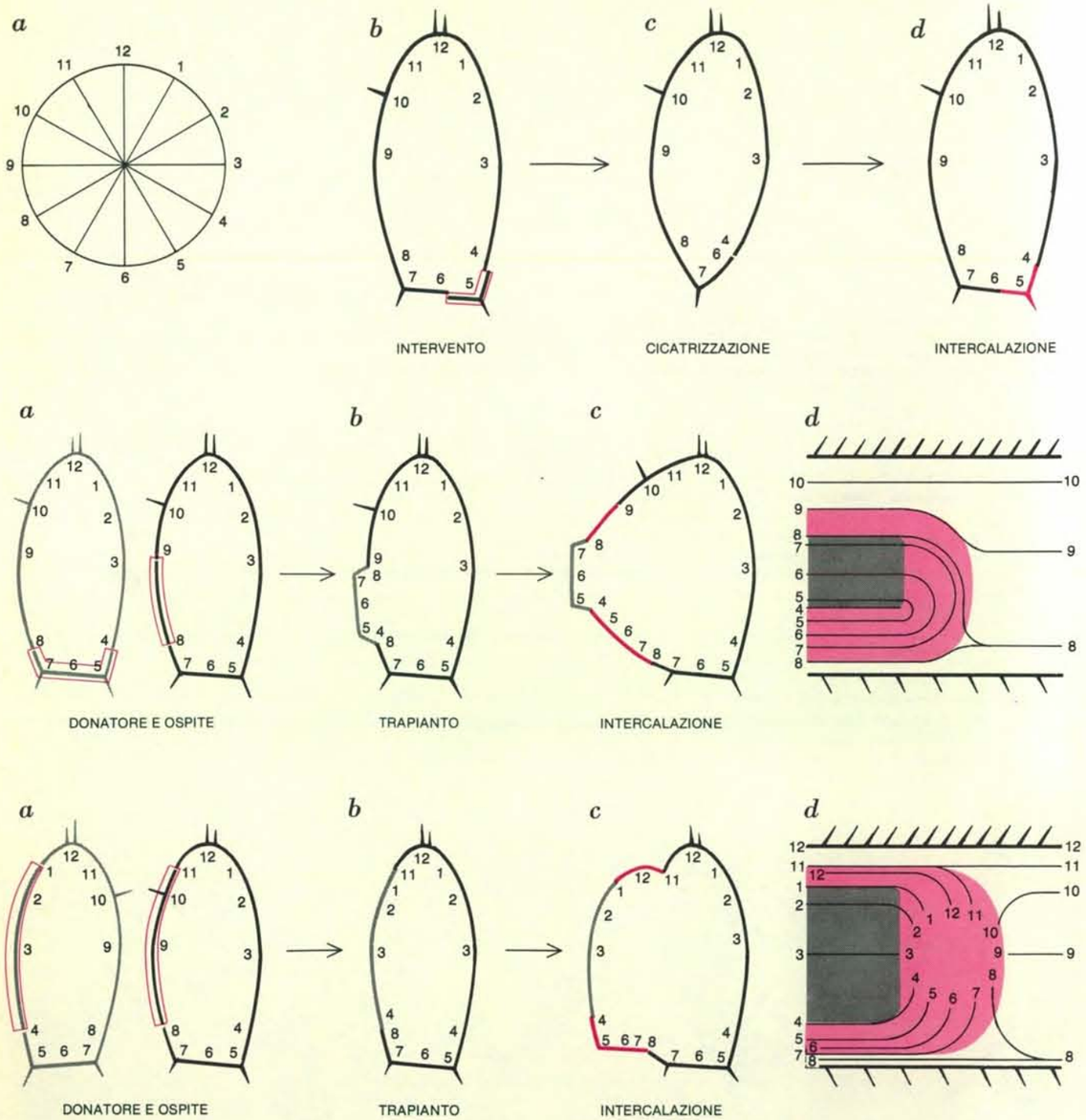




L'organizzazione d'una zampa di tritone, assomiglia a quella dei segmenti di zampa di blatta, lo schema dei livelli prossimali-distali viene applicato tuttavia all'intera zampa anziché a un solo segmento. Quando, come è indicato a sinistra, i livelli d'organizzazione intermedi vengono asportati e la parte distale viene trapiantata a metà di quella prossimale, mettendo a contatto la porzione di livello *E* con il moncherino di livello *A*, la rigenerazione intercalare nel punto d'innesto riproduce i livelli mancanti. Quando invece, come a destra, una zampa amputata a livello prossimale viene impiantata su un moncherino tagliato a livello distale, non viene interposto il livello tra *B* e *D* come avviene nella tibia di blatta.

lare, lunga circa la metà di un segmento normale.

Uno degli autori di questo articolo (Susan Bryant), lavorando in collaborazione con Laurie E. Iten e David L. Stocum, ha effettuato esperimenti simili sulle zampe di tritoni e salamandre, trapiantando blastemi di rigenerazione formatisi dopo l'amputazione a un livello prossimale-distale, su un moncherino di un ospite amputato a un livello diverso. Si è scoperto che trapiantando un blastema proveniente dalla parte anteriore dell'arto su un moncherino della parte posteriore si induce la crescita o la rigenerazione intercalare del livello intermedio mancante. Per ragioni ancora sconosciute, il reciproco trapianto di un blastema ottenuto dalla parte posteriore di un arto su un moncherino della parte anteriore non provoca rigenerazione intercalare:



Negli esperimenti schematizzati in questa figura si rivela l'organizzazione d'una zampa di blatta intorno a un secondo asse meridiano; i numeri segnati lungo la circonferenza servono a identificare le zone d'organizzazione. I valori di posizione vengono interposti per rigenerazione circolare. In alto, in seguito ad asportazione chirurgica della zona d'organizzazione 5 (*b, in colore*) le cellule delle zone adiacenti 4 e 6 cicatrizzano insieme (*c*). Nel giro di due mute la rigenerazione intercalare (*d*) permette di rimettere al suo posto, con interposizione di nuovo tessuto, la zona estirpata, con la sua caratteristica fila di setole. Al centro, quando viene eseguito un trapianto di donatore con valori di posizione da 4 a 8, su un ospite in posizione compresa tra i valori 8 e 9 (*a, b*), si verifica una rigenerazione intercalare appropriata su entrambi i bordi del trapianto (*c*); la figura *d* rappresenta l'intercalazione di tessuto rigenerato vista lateralmente. In basso, il trapianto dal donatore all'ospite (*a, b*) sostituisce i valori di posizione circolari compresi tra 8 e 11 con valori compresi tra 1 e 4. Su entrambi i lembi del trapianto si ha una crescita per interposizione (*c*) che dà origine a una circonferenza completa all'estremità del trapianto; la circonferenza distale è rappresentata schematicamente in *d*. Una circonferenza intercalata è equivalente a una formata per amputazione e origina la medesima reazione rigenerativa, cioè la produzione delle strutture della zampa che si trovano in posizione distale alla circonferenza.

I SEGRETI DI UN BUON VINO

Nel Castello di Brolio, arroccato tra boschi rocce e vigneti, nella regione dell'Alto Chianti, la Famiglia Ricasoli produce vino da più di 8 secoli. A Bettino Ricasoli (1809-1880), «il barone di ferro», Primo Ministro del Regno d'Italia dopo Cavour, si deve la definizione delle proporzioni di uve per ottenere il miglior vino del Chianti: sangiovese 70%, malvasia e trebbiano 15%, canaiolo nero 15%. Oggi questo vino si chiama **Chianti Classico;** la sua origine è protetta dallo Stato con la legge DOC = Denominazione di Origine Controllata; si differenzia dal vino nel resto della Toscana collinare, chiamato semplicemente **Chianti.**

Qualità, prestigio, tradizione contraddistinguono i vini Ricasoli, vanto della enologia italiana e mondiale. Eccone una scelta ed una piccola guida per gustarli meglio:

Castello di Brolio: famoso **Chianti Classico,** invecchiato per non meno di tre anni in fusti di rovere, si accompagna al meglio con tutte le carni saporite, in particolare arrosti e cacciagione. Va stappato qualche ora prima e servito a 16°-18° di temperatura.

Brolio Riserva: frutto di annate particolarmente felici di Brolio, invecchiato in rovere per 5 anni ed oltre, è vino da grandi occasioni. Si abbina ai piatti di carne di maggiore impegno ed ai formaggi stagionati. Va aperto e ambientato molte ore prima di essere bevuto a non meno di 18°.

Gaiole Chianti Classico: di invecchiamento in rovere e gradazione minore del Brolio, prende nome dal comune chiantigiano in cui si trova il Castello di Brolio. Si adatta pienamente alle carni rosse ed agli arrosti. Va servito a temperatura ambiente, cioè a 16°-18°.

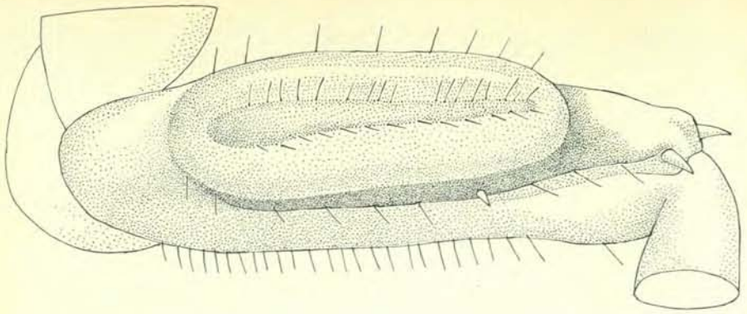
Brolio Bianco: prodotto con le uve di Pinot dei vigneti di Brolio, ha delicatezza e profumo ed è ideale con antipasti e crostacei. Si serve sempre fresco, a 8°-10°.

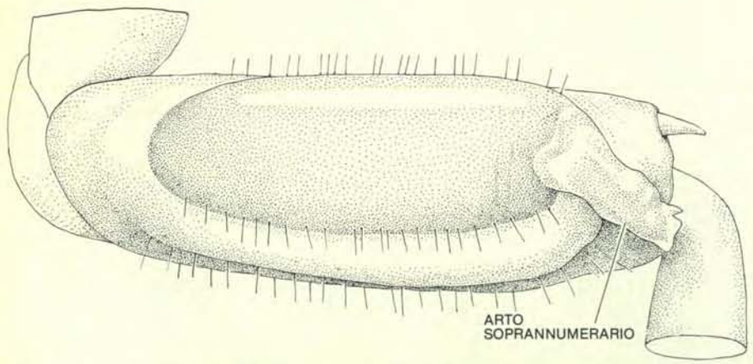
Brolio Rosè: morbido e profumato, accompagna tutti i piatti delicati. Si beve fresco a 10°.

Brolio Vin Santo: di antica origine, di produzione artigianale, occorrono 3/4 anni per vinificarlo e altri 3/4 di invecchiamento in piccole botti per ottenerlo. Nobilita tutti i dessert, e va bevuto ghiacciato (6°-8°) in piccoli bicchieri.

Albia Bianco: vino bianco dal gusto deciso, di sapore pieno, ideale con i piatti di pesce, è ottimo con gli antipasti ed i primi piatti. Va servito sempre fresco, a 8°-10°.

Chianti Ricasoli: meno impegnativo del Chianti Classico, asciutto, è vino da tutto pasto, ideale con i primi piatti saporiti e le carni in genere. Si beve a temperatura ambiente, intorno ai 16°.





In questi disegni appaiono le zone di crescita intercalare sulla circonferenza della zampa di una blatta, in riferimento alle illustrazioni a pagina 48 in cui sono raffigurate in modo schematico zampe viste lateralmente. Nella figura in alto si può notare una rigenerazione intercalare su entrambi i bordi del trapianto. In basso, dove la crescita intercalare si è risolta nella formazione d'un'intera circonferenza, si è sviluppato un arto soprannumerario distale al femore.

l'arto risultante mostra una struttura discontinua.

Tutti questi esperimenti indicano che le cellule dei segmenti di zampa di blatta nei differenti livelli prossimali e distali e la zampa in corso di rigenerazione d'un anfibio possiedono informazioni diverse, legate alla posizione. Inoltre l'interazione di cellule aventi tali diversi valori riguardanti la posizione porta a una stimolazione locale della crescita e all'interposizione di valori intermedi all'interno del nuovo tessuto. Per proseguire gli esperimenti è necessario effettuare innesti che mettano a contatto cellule provenienti dal medesimo livello in senso prossimale o distale, ma da posizioni diverse in senso circolare.

Interazioni trasversali

Uno di noi (French) ha condotto questo tipo di esperimenti con zampe di blatta, collaborando all'inizio con Bullière. Per esempio, la rimozione di una stretta striscia longitudinale di tegumento (epidermide e cuticola) del femore di una blatta viene seguita dalla cicatrizzazione della ferita e dall'avvicinarsi di due gruppi di cellule epidermiche che si trovano originariamente a una certa distanza l'uno dall'altro. L'interazione stimola una crescita localizzata e una rigenerazione intercalare; la parte di struttura intermedia mancante viene ricostituita e la zampa recupera la sua normale circonferenza.

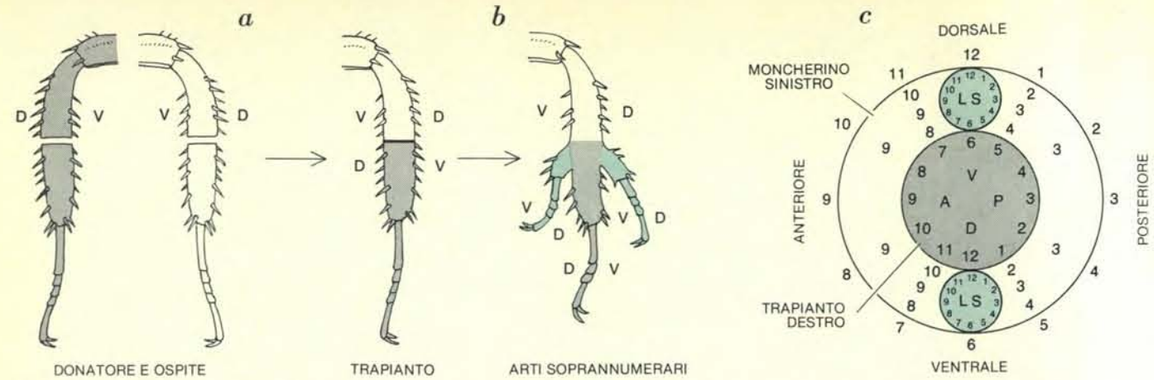
Una variante di questo esperimento consiste nel trapiantare una porzione rettangolare di tegumento dalla zampa di un donatore su una posizione anomala (lungo la circonferenza) della zampa di un ospite, senza alterare la posizione nel senso prossimale-distale. Vengono di nuovo messe a contatto cellule epidermiche provenienti da differenti posizioni stavolta lungo entrambi i bordi dell'innesto: questo stimola una crescita localizzata nei punti di congiunzione. Le strutture che normalmente stanno tra la parte ospite e la parte trapiantata divengono intercalate nel nuovo tessuto. Le strutture intercalate, comunque, sono solo quelle che normalmente sono localizzate sulla più breve delle due vie possibili lungo la circonferenza della zampa. Se in un trapianto di questo tipo si trovano di fronte cellule di donatore e di ospite provenienti da posizioni diametralmente opposte lungo la circonferenza della zampa, la successiva intercalazione di tessuto può formare l'una o l'altra delle semicirconferenze intermedie.

Si ottengono risultati simili anche quando le strisce di tessuto trapiantato vengono collocate su differenti segmenti delle zampe di una blatta o su differenti livelli prossimali-distali del medesimo segmento di una zampa. Questo fa pensare che le cellule attribuiscono differenti valori relativi alla posizione in rapporto alla collocazione in senso circolare, indipendentemente dal livello prossimale-distale. Come con i trapianti che riguardano l'asse prossimale-distale del segmento di zampa, l'interazione di cellule con differenti valori relativi alla posizione lungo la circonferenza del segmento della zampa porta a una stimolazione locale di crescita e all'intercalazione di valori intermedi.

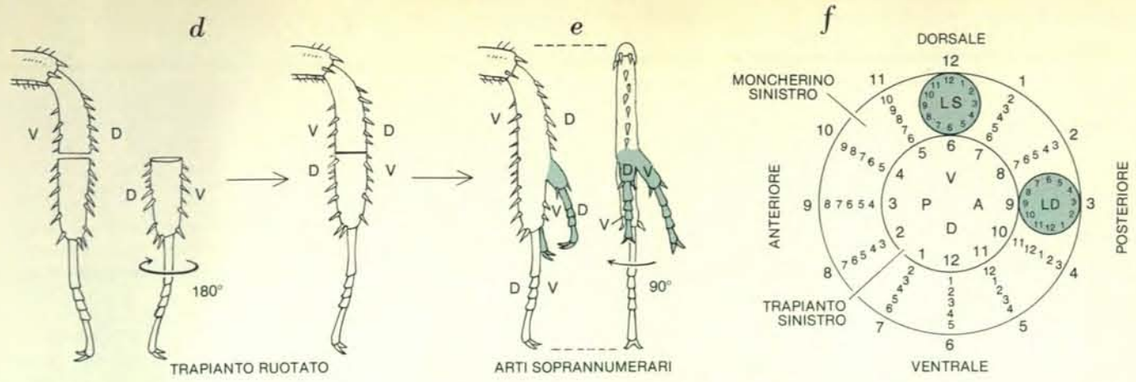
Strutture soprannumerarie

Molte operazioni di trapianto su zampe di blatta e tritoni provocano bizzarre rigenerazioni multiple. Uno studio di queste strutture soprannumerarie ci fornisce ulteriori informazioni sulle leggi che regolano le interazioni cellulari nelle zampe. Ricerche in questa direzione sono state effettuate sulle blatte da Bohn, Bullière e uno di noi (French) e sui tritoni da Iten e da Susan Bryant. Per esempio, quando una striscia longitudinale di tegumento di zampa di blatta viene trapiantata su un ospite in una posizione anomala lungo la circonferenza, si forma una zampa soprannumeraria a ciascuna estremità del trapianto in alcune combinazioni ma non in altre. Analizzando questi risultati si può dimostrare che le zampe soprannumerarie vengono prodotte solo quando l'estremità del tessuto trapiantato, il tessuto ospite adiacente e il nuovo tessuto intercalato fra quello dell'ospite e del trapianto, formano una circonferenza completa. Questa osservazione suggerisce una conclusione importante: le strutture distali si rigenerano non solo nel punto dove è stata compiuta l'amputazione, ma anche in tutte le altre zone in cui si possa formare una circonferenza completa. Inoltre, gli esperimenti confermano che l'esistenza di una circonferenza completa è il requisito fondamentale per una rigenerazione distale.

La rigenerazione a partire da una circonferenza completa spiega le zampe soprannumerarie che appaiono dopo altri tipi di trapianto eseguiti su zampe di blatta e di tritone. Quando si trapianta una zampa sinistra amputata di blatta o il blastema di rigenerazione proveniente dal moncherino di una zampa sinistra di tritone sul moncherino di una zampa destra, lo scambio da sinistra a destra capovolge uno degli assi trasversali del trapianto rispetto all'asse trasversale del moncherino. Si è scoperto che dopo un certo periodo di tempo dalla giunzione tra il moncherino e il trapianto sorgono due serie di parti distali soprannumerarie. Se è l'asse anteroposteriore quello capovolto nel trapianto, allora le parti soprannumerarie di zampa spuntano a monte e a valle della giunzione; se invece è stato capovolto l'asse dorsoventrale, le parti soprannumerarie di zampa si svi-



In molte blatte, quando gli esperimenti di trapianto portano le cellule a un contatto anormale, appaiono zampe soprannumerarie. Nell'esperimento a sinistra la metà distale di una tibia destra del donatore è stata trapiantata sul moncherino della metà prossimale d'una tibia sinistra dell'ospite (*a*). Questo trapianto mette le cellule dell'ospite e del donatore con valori diversi in senso circolare in opposizione le une rispetto alle altre in tutti i punti, salvo due. Si formano allora due zampe soprannumerarie (*b*), nei due punti di massima incongruenza. Le zampe soprannumerarie sono immagini speculari del tarso del donatore e si formano perché la crescita intercalare nel punto di



giunzione del trapianto ha originato circonferenze complete nelle posizioni di massima incongruenza tra le cellule dell'ospite e del tessuto trapiantato (*c, in colore*). La direzione dell'intercalazione più breve su ciascun lato di queste posizioni determina l'ubicazione dei valori di posizione e quindi il senso (destro-sinistro) e l'orientamento degli arti soprannumerari prodotti. Nell'esperimento a destra la tibia amputata viene ruotata di 180 gradi e trapiantata (*d*). Due segmenti soprannumerari sono apparsi in questo caso, uno lateralmente e uno posteriormente (*e*). Per il fenomeno dell'intercalazione (*f*) si sono formate circonferenze complete in ciascuno dei due punti di rigenerazione.

luppano lungo i lati dorsale e ventrale della giunzione. Una caratteristica di queste parti soprannumerarie è che esse rigenerano in ogni caso mentre la tendenza a essere destre o sinistre può essere determinata in quanto esse conservano il medesimo orientamento del moncherino. In questo tipo di trapianto le cellule del moncherino e quelle del trapianto vengono poste le une accanto alle altre in tutti i punti intorno alla giunzione trapianto-ospite: ma solo in due zone le cellule del moncherino e quelle del trapianto hanno valori relativi alla posizione corrispondenti. In ogni altro punto le cellule differiscono nei loro valori relativi alla posizione e la loro interazione provoca l'inserimento di cellule che possiedono la serie più breve di valori intermedi. Come risultato si ottiene la formazione di una circonferenza completa in ciascuno dei due punti di massima incongruenza tra moncone e trapianto. La rigenerazione da queste circonferenze complete provoca la formazione di arti soprannumerari nelle posizioni considerate e con l'orientamento e il senso osservati.

Quando una zampa amputata di blatta viene trapiantata sul suo stesso moncherino dopo essere stata ruotata di 180 gradi non si formano zampe soprannumerarie oppure se ne formano due in varie posizioni intorno alla circonferenza di congiunzione tra moncherino e innesto e con diverso orientamento. Quando si può determinare il senso, una delle zampe soprannumerarie è una zampa sinistra, l'altra una destra. Se la zampa è ruotata di soli 90 gradi prima di essere trapiantata sul moncherino, non si forma alcun arto soprannumerario.

Un trapianto analogo, ruotato di 180 gradi, di blastema d'arto di tritone provoca la formazione di una zampa sinistra e di una destra soprannumerarie. Diversamente dalle blatte, comunque, la collocazione di queste zampe soprannumerarie è costante. Questi risultati possono venire spiegati dai medesimi principi che regolano l'intercalazione di tessuto e la rigenerazione distale a partire da una circonferenza completa, ammettendo che la distribuzione dei valori relativi alla posizione intorno alla circonferenza delle zampe di blatta sia regolare mentre la distribuzione di tali valori intorno alla circonferenza della zampa di tritone sia leggermente irregolare.

Il modello a coordinate polari

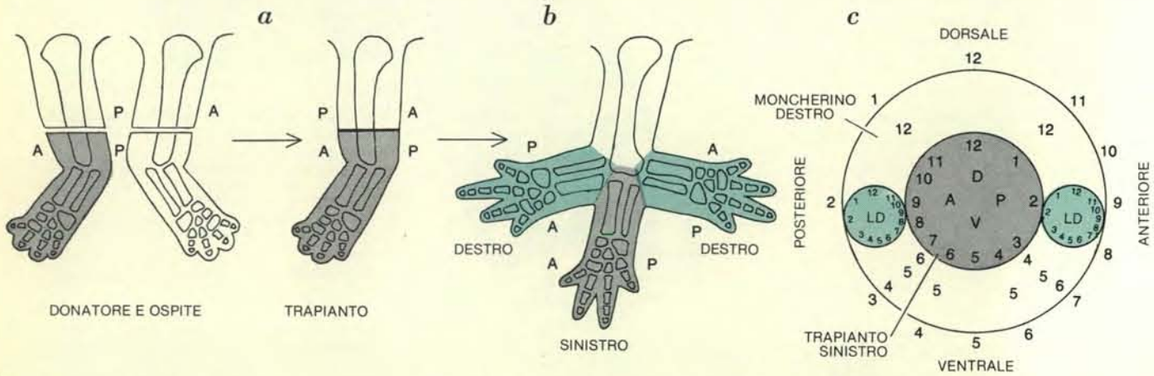
I dati sulla rigenerazione delle zampe mostrano chiaramente che, nonostante le differenze anatomiche tra le zampe di blatta e quelle di tritone, i due territori epimorfici obbediscono alle medesime regole fondamentali di interazione cellulare. Tali regole possono esser riassunte sotto forma di un modello semplice, che illustreremo qui di seguito. La zampa può venire rappresentata come un cono (o, a due dimensioni, come un disco). L'apice del cono (o il centro del disco) rappresenta l'estremità distale della zampa; la base del cono (o la circonferenza del disco) rappresenta il confine prossimale del territorio (o campo) epimorfico. Le due componenti separate delle informazioni circa la posizione spaziale, già sperimentalmente identificate, corrispondono pertanto a un sistema a coordinate polari per specificare la posizione nel campo. Ciascuna cellula del campo ha un singolo valore di posizione, caratterizzato da due componenti. Ciascuna componente definisce la posizione in senso circolare della cellula, in termini di 12 coordinate meridiane numerate in senso orario da 1 a 12. L'altra componente definisce il livello prossimale-distale della cellula con coordinate radiali indicate con lettere da *A* e *E*. Due coordinate sono sufficienti a determinare la posizione di ogni cellula.

Il comportamento delle cellule in un territorio è regolato da due leggi. La prima legge, che chiamiamo la legge dell'intercalazione più breve, stabilisce che se vengono messe a contatto cellule con valori relativi alla posizione normalmente non adiacenti, l'incongruenza induce una crescita nel punto della giunzione che continua finché non sono state intercalate cellule con valori di posizione intermedi. Come corollario di questa legge, si ricava che quando il contatto avviene lungo la circonferenza è sempre la serie più breve di valori intermedi delle due possibili serie a essere intercalata: per esempio, quando due valori, *3* e *6*, si trovano trapiantati a fianco, saranno interposti i due intermedi *4* e *5*, invece degli otto valori intermedi che separano il *6* dal *3* nell'altra direzione.

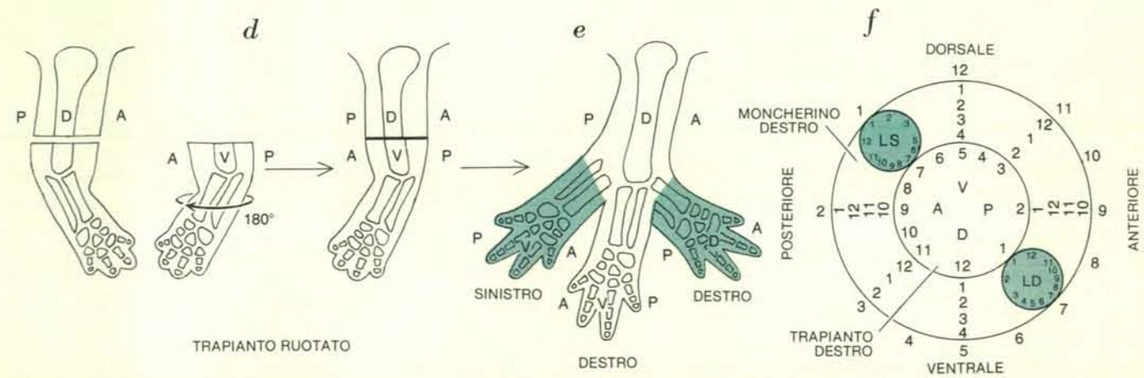
La seconda legge, che chiameremo legge della circonferenza completa per la rigenerazione distale, stabilisce che le cellule con valori di posizione equivalenti a un qualsiasi livello prossimale o distale possono originare nuove cellule con tutti i valori di posizione distali rispetto al proprio valore. Questo comunque può verificarsi esclusivamente quando sia esposta una sequenza completa di valori di posizione relativi alla circonferenza, sia per amputazione sia per intercalazione.

Il nostro modello, con le sue due semplici leggi, è d'aiuto per comprendere i complessi fenomeni di regolazione che si possono osservare nei territori secondari stabilizzati di due diversissimi organismi. Certamente le zampe di blatte e tritoni sono strutture tridimensionali mentre questo modello è solamente bidimensionale, perché trascura la componente interna-esterna. Nelle zampe di blatta, tuttavia, la struttura cuticolare è prodotta dalla secrezione di un'unico strato di cellule epidermiche e per i tessuti interni non vi sono indicazioni su alcune informazione relative alla posizione nei riguardi della rigenerazione. È possibile che la struttura dei muscoli sia derivata da una struttura primitiva nell'epidermide: il modello bidimensionale è perciò in grado di descrivere la zampa della blatta.

La zampa degli anfibi, più complessa, ha una parte centrale ossea circondata da muscoli e coperta da derma ed epidermide. Gli esperimenti indicano che



Anche nel tritone appaiono zampe soprannumerarie per reazione a trapianti che portano a contatti cellulari anomali. Nell'esperimento illustrato a sinistra, una zampa sinistra rigenerata, proveniente da un livello prossimale, viene trapiantata sul moncherino di una zampa destra, capovolgendo così l'orientamento anteroposteriore (*a*). La conseguenza, come nel trapianto analogo operato sulla blatta, è lo sviluppo di due arti soprannumerari, che sono immagini speculari dell'arto trapiantato (*b*). Le circonferenze intercalate (*c, in colore*) si



trovano nei punti di massima incongruenza. Nell'esperimento successivo (*a destra*) la rotazione di 180 gradi e il reimpianto di una zampa rigenerata (*d*) possono provocare la crescita di una o due zampe soprannumerarie in zone prevedibili (*e*). Se la rigenerazione appare nel quadrante dorsale, si tratterà d'una zampa sinistra, con la superficie ventrale dell'arto verso l'alto; se appare nel quadrante ventrale, sarà una zampa destra a orientamento normale. Anche in questo caso la rigenerazione è un prodotto di circonferenze complete (*f, in colore*).

anche qui le informazioni di posizione possono essere limitate a un modello bidimensionale. Gli esperimenti condotti da Bruce M. Carlson dell'Università del Michigan hanno dimostrato che nelle salamandre la manipolazione di ossa o epidermide dopo l'amputazione di una zampa non porta alla rigenerazione di zampe multiple, mentre la manipolazione di muscoli o derma può provocare tale fenomeno. Questo fatto fa pensare che solo muscoli e derma portino informazioni di posizione. Non ci sembra perciò troppo semplicistica l'ipotesi che i valori di posizione nelle zampe degli anfibi siano disposti in un cono, con il tessuto all'esterno e all'interno che è privo di informazioni di posizione, ma che è in grado di realizzare la propria struttura in base al modello primario contenuto nel derma e nel muscolo.

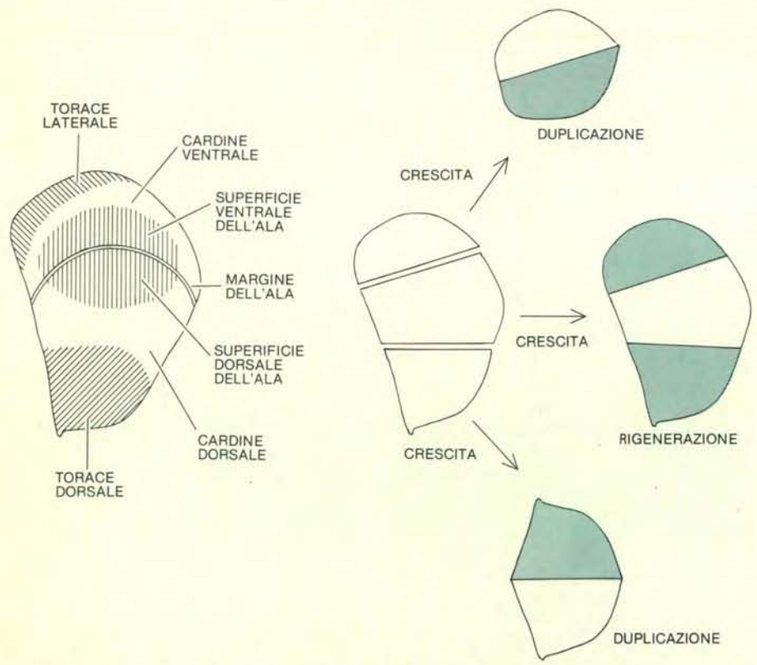
Regolazione del disco imaginale

Un'ipotesi esposta da Wolpert è divenuta praticamente un articolo di fede per molti biologi studiosi dello sviluppo. Secondo questo ricercatore parecchi organismi possiedono lo stesso meccanismo di base per determinare l'informazione sulla posizione e le differenze tra gli organismi nascono dal fatto che le loro cellule interpretano in modi diversi l'informazione sulla loro posizione quando si differenziano. Questo principio dell'universalità dell'informazione di posizione è stato suggerito in parte dalla scoperta dell'universalità in altri meccanismi biologici fondamentali, in parte dai risultati degli esperimenti di trapianto tra specie affini. Esso è stato oggi ulteriormente rafforzato dalla nostra recente scoperta che il modello a coordinate polari può spiegare non solo le capacità regolatrici delle zampe di organismi diversi, come anfibi e blatte, ma anche il comportamento regolatore mostrato da tipi completamente differenti di territori secondari: i dischi imaginali del moscerino della frutta, *Drosophila.* Questo insetto, l'animale da laboratorio preferito dai genetisti, fu adoperato per la prima volta da Thomas Hunt Morgan perché si prestava in modo eccellente alle manipolazioni genetiche, in un periodo in cui gli interessi di questo studioso stavano cambiando dal campo della rigenerazione e dell'embriologia a quello della genetica.

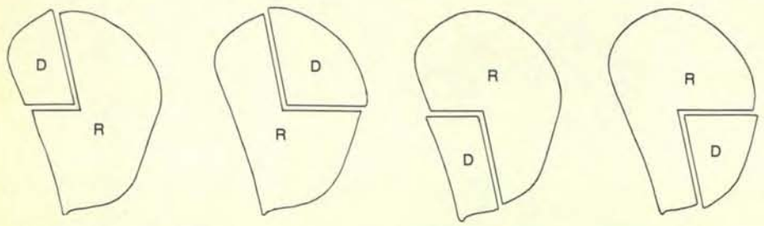
Lo stadio larvale immaturo della drosofila non presenta appendici: ali, antenne, zampe, bilancieri, le parti associate della testa e del torace e inoltre gli organi genitali del futuro adulto sono tutti rappresentati all'interno della larva in sviluppo come porzioni di tessuto in crescita, ma ancora indifferenziato, conosciute col nome di dischi imaginali. Tali dischi assumono la loro funzione specifica, cioè producono le varie parti dell'insetto adulto, solo dopo che la larva ha subito la sua caratteristica metamorfosi; secernono le strutture superficiali ricoperte di cuticola, come setole, peli e organi di senso secondo schemi complicati ma riproducibili.

I dischi imaginali della drosofila sono più piccoli delle zampe di una blatta o di un tritone di un'intero ordine di grandezza, ma tuttavia possono essere sottoposti a vari interventi di microchirurgia che possono mettere in evidenza le loro capacità regolatrici: per esempio, quando un pezzo di disco imaginale di zampa viene trapiantato nell'addome di una larva ospite della stessa età, subisce metamorfosi assieme all'ospite e si differenzia in parti riconoscibili di una zampa. Gerald A. Schubiger, dell'Università di Washington, ha dimostrato che le strutture specifiche formate dipendono dalla parte del disco trapiantata: questo gli ha permesso di ricostruire una completa mappa «del destino» del disco imaginale della zampa, Mappe analoghe sono state stilate per tutti gli altri dischi principali, per esempio per il disco alare.

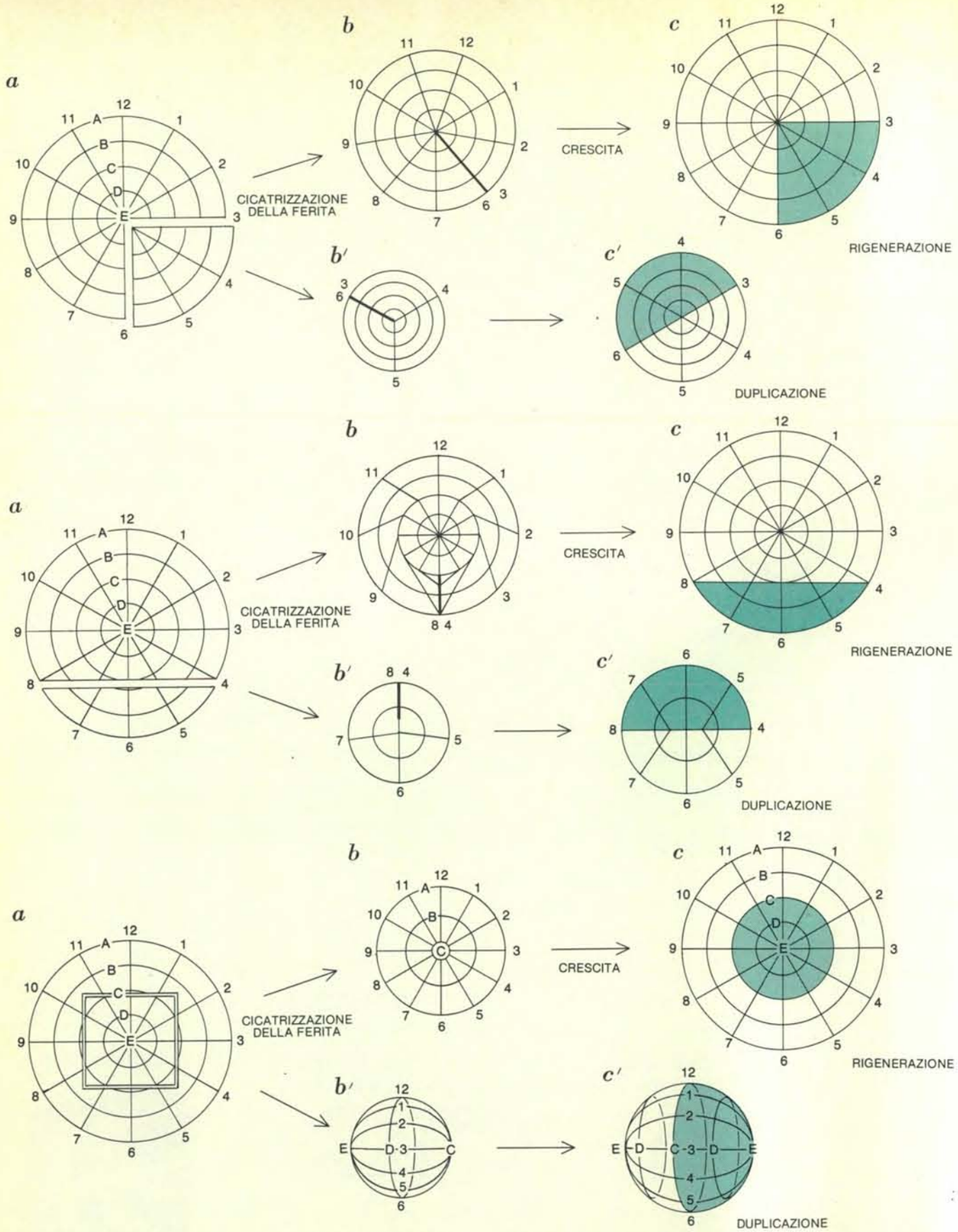
Effettuando questo tipo di esperimenti si è scoperto che parti complementari dei dischi imaginali originano parti complementari delle strutture dell'animale adulto: in altri termini, non esiste una regolazione dello sviluppo quando i frammenti del disco sono soggetti a un'immediata metamorfosi in una larva ospite. Probabilmente il motivo è dovuto al fatto che i dischi imaginali sono campi epimorfici e quindi i frammenti sono incapaci di regolarsi, a meno che sia loro fornita la possibilità di crescere prima della metamorfosi. Se tuttavia un frammento di disco viene trapiantato in un ospite adulto, invece che in una larva, le condizioni ormonali gli permettono di



Anche i dischi imaginali della larva del moscerino della frutta, *Drosophila,* possiedono capacità regolatrici. A sinistra si vede il disco imaginale che darà origine a un'ala quando la larva subirà la metamorfosi: le scritte nell'illustrazione indicano il destino normale delle sue diverse parti. Quando questo disco viene tagliato in tre pezzi (*al centro*) ed essi sono coltivati e trasferiti su tre diverse larve ospiti per seguirne la metamorfosi, il pezzo centrale si rigenera verso l'esterno in entrambe le direzioni (*in colore*), mentre i pezzi laterali si limitano a duplicarsi.



Tra i vari esperimenti di rigenerazione del disco imaginale, sono state eseguite queste quattro prove di divisione in due settori, l'uno di 270 gradi e l'altro di 90 gradi. Qualunque sia stato il modo di tagliare il disco, dopo la coltura e il trasferimento, il frammento di maggiori dimensioni ha rigenerato la parte mancante, mentre quello più piccolo si è limitato a duplicare se stesso.



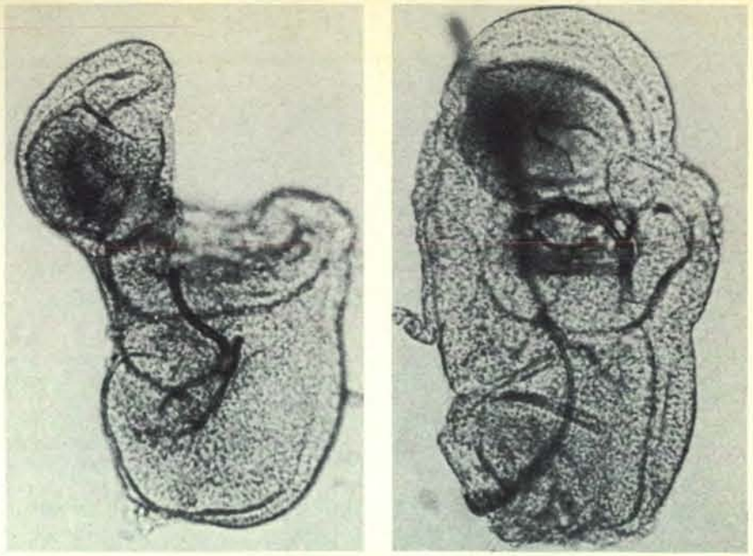
Ai risultati ottenuti coltivando frammenti di dischi imaginali dell'ala si applicano le leggi di interazione di un modello a coordinate polari. Nelle figure in alto vengono coltivati settori di 270 e di 90 gradi (*a*). Nei frammenti i margini tagliati si riattaccano cicatrizzandosi (*b, b'*), per cui, in entrambi i casi, vengono messe a contatto le posizioni 3 e 6. Si applica allora la legge dell'intercalazione più breve e vengono inseriti i valori mancanti, 4 e 5 (*c, c', in colore*). Per il settore di 90 gradi ciò equivale alla duplicazione, mentre per quello di 270 gradi il risultato è la rigenerazione. Nelle figure al centro, il disco imaginale viene tagliato in due parti (*a*). Entrambi i frammenti cicatrizzano contraendosi lungo i bordi tagliati e portando a contatto le posizioni 8 e 4 (*b, b'*). La medesima legge conduce a una rigenerazione completa (*c*) e a una duplicazione (*c'*). Nell'esperimento in basso, vengono coltivati separatamente un frammento centrale del disco alare e la parte periferica (*a*): la trasformazione distale porta alla duplicazione del frammento centrale (*b', c' visto lateralmente*) e alla rigenerazione del frammento periferico (*b, c*). Su ciascun bordo del frammento è presente una seguenza circolare completa di valori di posizione.

crescere senza compiere metamorfosi, finché non viene ritrapiantato in una larva ospite. Quando i frammenti di disco vengono fatti crescere, in una coltura a breve termine, per una settimana o anche meno in un adulto, prima di essere ritrapiantati, mostrano un tipo di comportamento regolatore che all'inizio può parere insolito. Per esempio, Schubiger ha sperimentato che la metà mediale del disco della zampa può rigenerare la metà laterale mancante, ma quando la metà laterale è tenuta in coltura non è in grado di rigenerare la metà mediale, ma subisce una duplicazione, per cui si ottengono due copie delle strutture laterali della zampa.

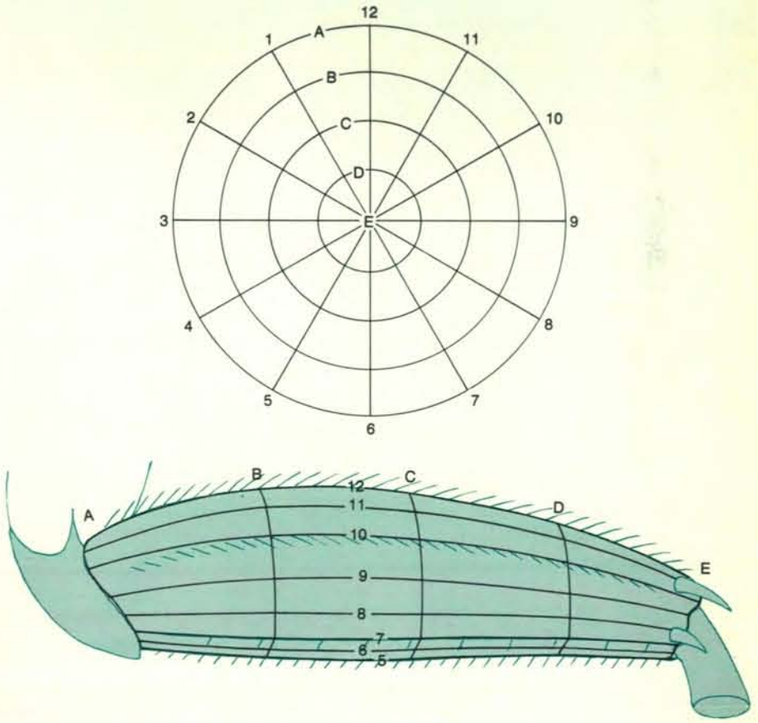
Molti frammenti di dischi imaginali di zampa, di ala e di altri organi sono stati sottoposti a esperimenti e si è potuto constatare una rigenerazione oppure una duplicazione. Inoltre i risultati sembrano rivelare l'esistenza di uno schema curioso: se un frammento di un disco imaginale ha la capacità di rigenerarsi, il frammento complementare di quel disco generalmente si duplica. La capacità rigenerativa dei dischi imaginali ci permette di considerarli come tipici territori secondari, ma fino a poco tempo fa le proprietà di questa duplicazione sembravano piuttosto misteriose.

Uno studio minuzioso della capacità regolatrice del disco dell'ala, effettuato da uno di noi (Peter Bryant) ha dimostrato che il suo comportamento (e probabilmente quello degli altri dischi imaginali) può essere facilmente spiegato dal modello a coordinate polari. Quando il disco viene tagliato in tre pezzi, quello centrale si rigenera verso l'esterno e i due pezzi marginali si duplicano. Se i frammenti vengono ottenuti tagliando lungo linee radiali partendo dal centro del disco e andando verso la periferia, essi pure si rigenerano o si duplicano. Si è scoperto che ciascuno dei quattro settori di 90 gradi del disco dell'ala si duplicano mentre i settori complementari di 270 gradi si rigenerano.

Questi risultati possono venire spiegati con la legge dell'«intercalazione più breve». Per poter far ricorso a questa legge occorre mettere a contatto cellule con diversi valori di posizione nel tessuto e ora appare chiaro che ciò avviene effettivamente in frammenti di disco imaginale, come conseguenza di una cicatrizzazione della ferita. I settori di 90 e 270 gradi si cicatrizzano fondendo i due margini tagliati perpendicolarmente, portando a contatto cellule con valori di posizione che distano di 90 gradi lungo la circonferenza del disco. L'interposizione più breve tra questi valori messi a contatto spiega in tal modo la rigenerazione osservata dei settori circolari di 270 gradi e la duplicazione di quelli di 90 gradi. Il medesimo ragionamento vale per la rigenerazione e la duplicazione in frammenti con i margini tagliati diritti, dove la cicatrizzazione della ferita provoca contrazione del lembo tagliato e contatto di valori di posizione lungo tale lembo. In generale con questo tipo di ragionamento si può



In queste microfotografie si può osservare la cicatrizzazione d'un frammento di disco imaginale dell'ala. Il processo inizia già il primo o il secondo giorno successivo al trapianto del frammento nell'addome di drosofila. Qui appare un settore di 270 gradi coi bordi tagliati (*a sinistra*) e dopo che i bordi si sono saldati (*a destra*) prima del fenomeno di rigenerazione intercalare.



In questa illustrazione il sistema a coordinate polari è rappresentato come un disco (*in alto*). I dodici meridiani sono identificati da numeri; i livelli radiali, centrali, periferici e intermedi, sono identificati da lettere in ordine alfabetico. Questo sistema permette di specificare l'informazione di posizione in un territorio (o campo) epimorfico e, assieme alle due regole enunciate dagli autori, è un modello schematico utilissimo per prevedere il comportamento delle cellule contenute in tale territorio. Come si vede dall'esempio del femore della blatta (*in basso*), questo modello rappresenta valori di posizione da quelli prossimali ai distali nelle sue divisioni dalla periferia al centro, mentre nelle sue divisioni in senso circolare rappresenta valori dorsoventrali e anteroposteriori. L'applicazione di questo schema di riferimento ai dischi imaginali è recente.

prevedere che i frammenti di grandezza maggiore alla metà del disco si rigenerano, mentre i pezzi più piccoli si duplicano. Questo fenomeno è proprio ciò che si osserva sperimentalmente.

Il modello a coordinate polari spiega anche il comportamento dei frammenti centrale e periferico del disco dell'ala, che sembrava essere in contraddizione col comportamento degli altri frammenti. Come abbiamo già visto i frammenti staccati con due tagli di solito rigeneravano verso l'esterno in entrambe le direzioni; questa regola valeva per tagli effettuati in parecchie direzioni: longitudinale, trasversale e diagonale. Un quadrato centrale dovrebbe perciò rigenerare tessuto in tutte le quattro direzioni, ma quando tale frammento è stato sottoposto a esperimento, si è visto che si duplicava senza rigenerare. Inoltre il rimanente frammento periferico ha potuto rigenerare il centro, secondo la legge che i pezzi complementari rigenerano e duplicano. Questo comportamento paradossale comunque è esattamente ciò che si deduce dalla legge della «circonferenza completa»: sia il quadrato centrale, sia il frammento periferico hanno un giro completo di valori di posizione esposti al margine di taglio, e pertanto ci si può aspettare in ogni caso una trasformazione distale. L'aggiunta di strutture destinate a divenire distali corrisponde alla rigenerazione per i frammenti periferici e alla duplicazione per quello centrale.

Recentemente uno di noi (Peter Bryant), in collaborazione con John L. Haynie, ha dimostrato direttamente che nel disco imaginale dell'ala può avvenire una rigenerazione intercalare. Abbiamo staccato due frammenti dei lembi opposti del disco dell'ala: uno è stato destinato a dar origine alla parte dorsale del torace e l'altro alla parte laterale: quando i due frammenti sono stati posti in coltura, da soli o mescolati meccanicamente con frammenti identici, si sono duplicati; quando invece i due pezzi delle estremità opposte del disco sono stati mescolati assieme, sono stati in grado di formare, grazie alla rigenerazione intercalare, le strutture intermedie mancanti.

Costituzione di territori secondari

Il nostro modello fornisce una spiegazione sul modo con cui si stabilizza inizialmente nell'embrione una struttura all'interno dei territori epimorfici secondari. Fra il 1910 e il 1930 Harrison e i suoi collaboratori dimostrarono che negli embrioni di anfibi la regione dove dovrebbero crescere le zampe si comporta in molti modi come un tipico territorio epimorfico. Questo appare particolarmente chiaro quando il territorio presunto della zampa di un embrione viene trapiantato secondo un qualsiasi orientamento all'interno della medesima zona di un altro embrione o su un fianco dell'embrione stesso dietro la zona in cui si presume si sviluppino le zampe anteriori. Oltre alla zampa che si sviluppa dall'innesto, possono generarsi una o due zampe soprannumerarie. Quando vennero effettuati per la prima volta questi trapianti ci si accorse che quasi tutti i turbamenti prodotti nel territorio dove si presumeva dovesse svilupparsi la zampa, potevano provocare la duplicazione. Ora comunque sappiamo che le zampe soprannumerarie si formano unicamente nelle posizioni dove sarebbero previste se il territorio presunto funzionasse come un territorio secondario stabilizzato, in accordo con i principi che abbiamo delineato.

Esaminando i primi stadi dello sviluppo si nota che il territorio suddetto di un giovane embrione possiede già una stabile organizzazione anteroposteriore. Questa si mantiene durante il successivo sviluppo anche se il territorio viene staccato e trapiantato in un embrione ospite secondo un orientamento anomalo. Comunque l'organizzazione dorsoventrale di questo territorio non è definitivamente stabilita fino a uno stadio molto più avanzato: immediatamente prima che la zampa inizi a spuntare dal corpo. Anche in altri territori secondari, come l'occhio e l'orecchio, si osserva una sequenza del tutto simile nello stabilirsi degli assi.

A prima vista la sequenza secondo cui si stabilisce un asse anteroposteriore e poi uno dorsoventrale suggerisce l'ipotesi che il territorio sia organizzato secondo un sistema di coordinate ortogonali per l'informazione di posizione. Tuttavia è possibile interpretare la sequenza secondo il nostro modello dei territori secondari, come cercheremo di spiegare qui di seguito. Supponiamo che i valori di posizione corrispondenti alla zona anteriore e posteriore della parte più prossimale del territorio siano i primi a stabilizzarsi e che i valori corrispondenti alle parti dorsale e ventrale delle regioni più prossimali si stabilizzino in uno stadio più avanzato. L'intercalazione secondo la via più breve fra i valori di posizione esistenti provoca quindi la formazione di una circonferenza completa nella zona più prossimale del territorio. In seguito, la trasformazione distale a partire dalla circonferenza prossimale stabilisce tutti i valori di posizione del territorio più distali.

Abbiamo osservato che in diversi organismi la rigenerazione delle strutture larvale e adulta, e forse anche la formazione e la crescita di queste strutture nei primi stadi embrionali, è regolata dalle medesime leggi fondamentali di comportamento cellulare. Per mezzo di semplici esperimenti chirurgici abbiamo identificato quelli che presumiamo siano i tipi di interazione cellulare che consentono lo sviluppo di forme e strutture nei territori secondari. Non siamo tuttavia ancora in grado di spiegare i meccanismi di queste interazioni cellulari e dobbiamo riconoscere di non sapere nulla riguardo la natura dell'informazione o dei segnali che vengono trasmessi tra le cellule di un organismo in sviluppo. La spiegazione di questi meccanismi di sviluppo, a nostro avviso, sarà d'importanza fondamentale per la biologia così come lo è stata la recente scoperta dei meccanismi genetici.

Le calcolatrici della Texas Instruments si trovano nei seguenti punti vendita qualificati:

PIEMONTE-VAL D'AOSTA. Ropolo s.a.s., Via Giannone 8a, Torino. **Legoval s.a.s.**, Via Torino 25, Aosta. **Electronics s.a.s. di E. Barale**, Via Statuto 10/B, Cuneo. **Emit**, C. so Montecucco 59, Torino. **Vel Foto**, Via Dossenam 27, Alessandria. **Lavander Sergio**, Corso Italia 67, Trino Vercellese. **Manassero**, P. Ellero 30, Mondovì. **Turletto Mario**, Via Sestriere 47, Moncalieri. **Nulla Sfugge**, P. C. Felice 23, Torino. **LOMBARDIA. Texas Instruments Italia S.p.A.**, Via Pattari 2, Milano. **Foto Zanardelli**, Corso Zanardelli 21, Brescia. **Guarnieri**, Corso Garibaldi 72, Cremona. **Bongasciù**, P. Quadrivio 1, Sondrio. **Malacreda**, Via Rusconi 14, Como. **Andreini**, Portici Sentierone 46, Bergamo. **Ropolo**, Viale Certosa 63, Milano. **Jenzi di T. Procacciante**, Passaggio Duomo 2, Milano. **Clup S.C.R.L.**, P. Leonardo Da Vinci 32, Milano. **Nuova Telecolor**, P.S.Fedele 2, Milano. **B.O.M.**, V.le Tunisia 50, Milano. **Elettrodomus di Aiselli**, Via Della Libertà 11, Pavia. **Casa Dell'Elettricità di Losi**, Via Umberto I, 68, Mantova. **Ceg**, Via R. Sanzio 29, Seregno (MI). **Beretta e Fioretti dei F. lli Moretto**, Via Italia 29, Monza. **Marco Pontiggia**, Corso Roma 3, Varese. **Ferrero Sergio**, Via Mameli 22, Busto Arsizio (Va). **F. lli Bresciani**, Via Casalino 8, Bergamo. **VENETO.**
Carlotto, Via Garofolino 22, Vicenza. **R.C.E. di P. Anzani**, Via Carlo Leoni 32, Padova. **Cartoleria Goliardica s.a.s. di F. Pasin**, Via VIII Febbraio 6, Padova. **Palesa**, Calmaggiore 10, Treviso. **Fanton Elettronica s.n.c.**, Via Venezia 106, Padova. Via Fusinato 34, Mestre. **C.I.M.E.**, Viale Verona 114, Vicenza. **Ing. Wladimiro Bedin**, Via XXX Aprile 40, Cornuda (TV). **Italtecnica di Adamo**, Via F. Marzollo 28/2°, Padova. **Berto Pasquale**, Via Dell'Officina 25, Cittadella (Pd). **Rag. E. Capovilla**, Gall. Trieste 8, Padova. **Vanotti S.p.A.**, Via Roma 14, Padova. **FRIULI, VENEZIA GIULIA. Mofert**, Viale Europa Unita 41, Udine. **Trevisan s.n.c.**, Piazza Duomo 3, Udine. **Centro Studi**, Piazza Cavedalis 3, Udine. **Fanton Udine s.r.l.**, Via E. Fermi 53, Feletto Umberto (UD). **Universaltecnica**, Corso U. Saba 18, Piazza Goldoni 1, Trieste. **Fototecnica Carducci**, Via Carducci 35, Trieste. **Brisotto Franco**, Corso Garibaldi 31, Pordenone. **Avanzo G. Succ.**, Corso Italia 17, Piazza di Cavana 7, Trieste. **TRENTINO, ALTO ADIGE. Fanton Bolzano s.r.l.**, Via Della Zecca 1, Bolzano. **Electronia S.p.A.**, Via Portici 1, Bolzano. **LIGURIA. Foto Bruno**, Piazza S. Giovanni 21, Imperia. **Ropolo s.a.s.**, Via Brigata Liguria 55/57/58, Genova. **Salviati Gianandrea**, Via I. Frugoni 9/R, Genova. **Foto Jolly di Gilardoni**, Via Boselli 3, Savona. **Foto Luciano di Romano**, Via Del Prione 299, La Spezia. **EMILIA, ROMAGNA. Borotti Mario**, Via XX Settembre 13, Piacenza. **Degano S.p.A.**, Via Cavallotti 49, Parma. **Minelli**, Via Bainsizza 7, Bologna. **Gambi**, Via Guidone 28, Ravenna. **Mari Mauro**, Corso Vittorio Emanuele 5, Modena. **Fratelli Fornasini s.n.c.**, Via Marconi 49, Bologna.
TOSCANA. Sare di Del Carlo, Via Vittorio Veneto 26, Lucca. **Paoletti Ferrero**, Via Il Prato 40/42/R, Firenze. **Petri Paris**, Via Fillungo 100, Lucca. **Elettrografica**, Via Curtatone Montanara, Pisa. **Fioresi Giuseppina Bianchi**, Piazza S. Francesco 57, Pistoia. **Pistoj**, Via Condotta, 22-67/R, Firenze. **Eraclex di Tafi Varis**, Via S. Ferrer 14, Piombino (LI). **Butali Benito**, Via Guelfa 22, Arezzo. **Casa Dello Sconto**, Via Toselli, Firenze. **Brogi Renato**, Piazza Gramsci 1, Siena. **Baini e Piccini**, Via Cavour 6/D, Carrara. **Orsini Mirella**, Piazza XX Settembre, Livorno. **Centro Casa**, Viale dei Mille, Firenze. **MARCHE. Eredi Rossi Giovanni**, Via Marconi 183, Ancona. **Cardinaletti Giorgio**, Corso Della Repubblica 59, Fabriano. **UMBRIA. Comer di L. Tattanelli**, Via Della Pallotta 20/D, Perugia. **LAZIO. Radiovittoria**, Via Luisa di Savoia 12, Via Ugo Ojetti 139, Via F. Gai 8, Roma. **Elettromercato Av**, Via Andrea Doria 34, Roma. **Di Veroli**, Piazza V. Emanuele II, 141/142, Roma. **All'Orologio**, Via Del Governo Vecchio 7, Roma. **Izzi Riccardo**, Via Nomentana 31/33, Roma. **Corsani**, Via Ottaviano 79, Roma. **Alessi**, Via Dei Serpenti 171, Roma. **Cartolibreria Nitti**, Via F. S. Nitti 66, Roma. **Datatron s.r.l.**, Via Mercadante 11, Roma. **FBM**, Via Flaminia 395, Roma. **De Angelis M. Pia**, Via Del Monte Oppio 14/16/18, Roma. Via Pietro Rosa 23, Lido di Ostia, Roma. **Eldo**, Viale Libia 42, Viale Marconi 158, Via Roberto Malatesta 249, Roma. **Union Shop**, Viale Regina Margherita 9, Roma. **Elettrostandar**, Via Furio Camillo 56, Roma. **CAMPANIA. Spot 2**, Via Roma 374, Napoli. **Aiello Salvatore**, Via Cervantes 37, Napoli. **Centro Sandoz di Norrito Enza**, Piazza Municipio 56, Napoli. **Casa Del Rasoio di Maiella Gaetano**, Piazza Garibaldi 75, Napoli. **Elvetico Mariano**, Via Scarlatti 128, Napoli. **Istituto Ottico Italiano Cav. Carlo La Barbera s.r.l. di Vittorio Ettore**, Via Roma 186/187, Napoli. **Electra dei F. lli S. & M. Petrosino**, Piazza Sedile Portanova 10, Salerno. **Giuseppe Maddaluna**, Via G. Mazzini 50, Caserta. **PUGLIA. Leocart**, Viale Unità D'Italia 63, Bari. **Favia S.M.C.**, Piazza Umberto 2, Bari. **Colucci Giuseppe**, Corso Umberto 89, Brindisi. **Decimo E di Maio**, Trav. XXXI C. Petroni, Bari. **CALABRIA. Angelo Randazzo S.p.A.**, Via Panebianco 220/240, Cosenza. **BASILICATA. Agesa**, Via Roma 143/145, Moliterno (PZ). **Giacumma Carlo**, Piazza Bonaventura 10, Potenza. **SICILIA. Angelo Randazzo S.p.A.**, Via Ruggero Settimo 51/53, Via G. B. Lulli 18, Palermo. Largo Dei Vespri 21, Catania. Via Ghibellina 32, Messina.

La solidificazione del cemento

Il processo che, con l'aggiunta d'acqua, trasforma la malta plastica in materiale compatto e resistente avviene tramite reazioni chimiche le quali danno luogo a un sistema di minuscoli grani compenetrantisi

di D.D. Double e A. Hellawell

Il cemento può essere descritto in modo semplice e schematico come un materiale calcareo presentato in commercio sotto forma di polvere sottile: con l'aggiunta di acqua, forma una malta plastica che si rapprende, cioè fa presa, e quindi indurisce fino a raggiungere la consistenza della pietra. Questo è pressappoco quanto viene dato in generale per scontato, ma non è forse altrettanto noto che lo sviluppo della resistenza meccanica è il risultato delle reazioni chimiche fra i componenti del cemento e l'acqua. Il processo di solidificazione si verifica per la formazione di una matrice rigida di interconnessione costituita dai prodotti di idratazione che sostituiscono a poco a poco l'acqua presente fra i granelli e rendono alla fine compatta la massa composita di cemento. L'idratazione implica alcuni processi molto complessi. Questo articolo ha lo scopo di descrivere alcune conoscenze recenti sull'argomento acquisite grazie agli studi condotti dagli autori presso l'Università di Oxford e da altri ricercatori.

Consideriamo le malte di calce - cioè miscele di calce viva (ossido di calcio, CaO) e sabbia - usate dai costruttori romani. Lo sviluppo della resistenza meccanica e la durata secolare di questi cementi non derivano tanto dalla reazione iniziale - scritta $CaO + H_2O \rightarrow Ca(OH)_2\downarrow$ per indicare che con l'aggiunta di acqua l'idrato precipita formando un colloide, ovvero un sistema nel quale le particelle di materiale finemente suddivise risultano disperse nell'acqua - quanto dal successivo lento processo di stagionatura nel corso del quale l'idrato colloidale si trasformava in carbonato di calcio per reazione con l'anidride carbonica atmosferica [$Ca(OH)_2 + CO_2 \rightarrow CaCO_3$]. Le ottime caratteristiche dei cementi romani, evidenti nelle numerose e stabili costruzioni che tuttora resistono, erano in gran parte dovute alla successiva scoperta che la calce miscelata con un materiale siliceo reattivo (laterizi frantumati o cenere vulcanica) forniva un cemento di maggiore resistenza alle sollecitazioni e all'azione dell'acqua. Il «cemento romano» realizzato in questo modo godeva di una grande rinomanza e conservò il favore dei costruttori fino alla fine del XVIII secolo con ben pochi miglioramenti o modifiche.

L'invenzione del cemento Portland malgrado alcune controversie sorte sulla sua origine è in generale attribuita a Joseph Aspdin, un imprenditore edile di Leeds, che nel 1824 prese un brevetto per «un cemento di qualità superiore somigliante alla pietra di Portland». Il suo cemento veniva preparato sinterizzando a temperatura elevatissima quantità in proporzioni fisse di calcare o di gesso (carbonato di calcio) e di argilla (silicati di alluminio). Il processo è alla base del metodo di fabbricazione ancora in uso. Durante la cottura i materiali grezzi si combinano per formare il cosiddetto clinker di cemento (sotto forma di pezzi tondeggianti che vengono poi macinati per ottenere il prodotto in polvere), contenente i silicati di calcio reattivi. I silicati conferiscono al cemento il carattere idraulico, cioè la proprietà di fare presa per reazione con l'acqua. Contrariamente ai vecchi leganti a base di calce, il cemento di Aspdin fa presa anche se immerso in acqua, alla cui azione è resistente per lungo tempo. Uno dei primi impieghi su vasta scala del cemento Portland si è avuta nella galleria del Tamigi, costruita fra il 1825 e il 1843 da Marc Isambard Brunel e da suo figlio Isambard Kingdom Brunel.

Il cemento Portland di tipo normale è ancora oggi di gran lunga il più importante dei cementi per quanto riguarda la quantità impiegata nelle costruzioni. Nella forma più diffusa viene mescolato con sostanze inerti (sabbia e altri materiali a granulometria controllata) per formare il calcestruzzo. Sono state realizzate numerose varianti al tipo fondamentale per soddisfare a requisiti particolari (come la presa rapida, il basso sviluppo di calore e la resistenza ai solfati), ma si tratta in ogni caso di modifiche relativamente secondarie nella composizione o nel metodo di preparazione.

Da questi prodotti si discosta in modo notevole il cemento alluminoso, anch'esso con proprietà idrauliche sebbene presenti composizione chimica e caratteristiche di idratazione del tutto differenti, ragioni per le quali costituisce un interessante materiale di paragone con il cemento Portland. Il cemento alluminoso, sviluppato in Francia e posto sul mercato dopo la prima guerra mondiale, viene ottenuto fondendo calcare e bauxite (idrossido di alluminio).

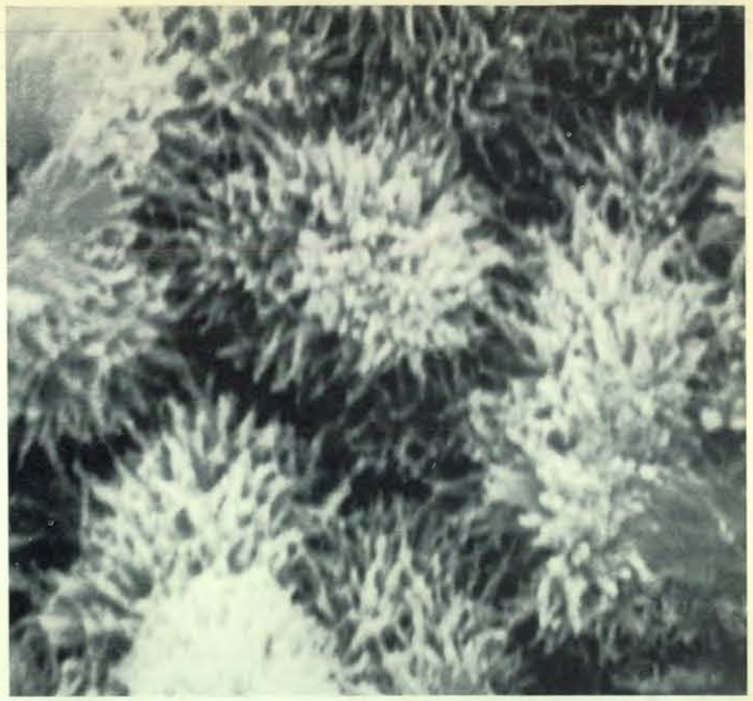
A differenza del cemento Portland, a cui si è più o meno pervenuti per tentativi ed errori, il cemento alluminoso è stato progettato per resistere in modo specifico ai solfati. Esso si è in effetti dimostrato adeguato per risolvere numerosi problemi costruttivi in condizioni ambientali nelle quali il cemento Portland non era adatto per la scarsa resistenza all'attacco chimico da parte dei solfati. Un altro vantaggio, ben presto riconosciuto e apprezzato, è il rapidissimo indurimento di questo particolare cemento. Quest'ultima caratteristica ne ha fatto una preziosa alternativa al cemento Portland, a più lunga stagionatura, in alcune applicazioni quali lavori di ripristino che debbano essere eseguiti rapidamente e la costruzione di elementi prefabbricati, dove un maggiore ritmo di produzione assicura un vantaggio economico che compensa in larga misura il costo relativamente maggiore.

L'esperienza acquisita nel tempo sui calcestruzzi a base di cemento alluminoso ha però dimostrato l'insorgere di fenomeni di degradazione a lungo termine accompagnati da diminuzione della resistenza e quindi l'impiego indiscriminato di questo materiale per qualunque condizione di sollecitazione viene oggi considerato assai discutibile. La maggior parte delle nazioni non consentono più l'impiego del cemento alluminoso nelle strutture portanti e in alcuni paesi sono stati imposti divieti o vincoli in tal senso. Molto spesso si tratta di provvedimenti presi in seguito a sinistri nei quali strutture portanti o perfino interi edifici costruiti con cemento alluminoso sono crollati, soprattutto per la degradazione del calcestruzzo.

Questo articolo descrive gli aspetti principali dell'idratazione e dello sviluppo della resistenza meccanica, confrontando il cemento Portland e quello alluminoso. Entrambi i cementi sono costituiti da un sistema ternario di ossidi di calcio, di silicio e di alluminio (CaO—SiO_2—Al_2O_3). I cementi commerciali presentano una composizione variabile entro limiti ristretti e contengono diverse impurità, costituite soprattutto da ossidi di ferro (meno del 10 per cento nel cemento Portland e fino al 20 per cento nel cemento alluminoso) e in misura minore da calce libera e da altri ossidi. Le impurità non hanno tuttavia alcun effetto sostanziale sulle proprietà di base dei cementi.

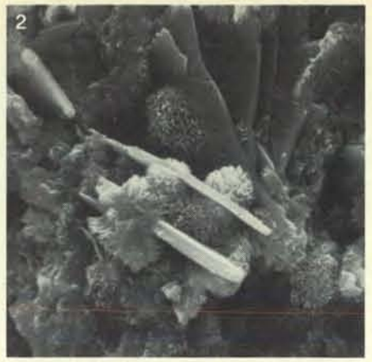
Il cemento Portland è fondamentalmente una miscela di silicati di calcio, in prevalenza silicato tricalcico e bicalcico, che si possono scrivere in forma abbreviata C_3S e C_2S (verrà qui impiegata una notazione semplificata nella quale gli ossidi di calcio, silicio, alluminio e ferro sono rappresentati rispettivamente da C, S, A e F, mentre l'acqua è indicata con H: così per esempio il silicato bicalcico che viene normalmente scritto $2CaO{\cdot}SiO_2$, nel sistema semplificato viene espresso con C_2S). Sono inoltre presenti percentuali minori di alluminio tricalcico (C_3A) e di una alluminoferrite di calcio avente una formula prossima a C_4AF. Le proporzioni tipiche in peso dei vari componenti sono in un normale cemento Portland: C_3S, 50 per cento; C_2S, 25 per cento; C_3A, 10 per cento: C_4AF, 10 per cento; altri ossidi, 5 per cento. Il cemento alluminoso, come indica il nome, contiene in prevalenza alluminato monocalcico ($CaO{\cdot}Al_2O_3$ scritto CA) accompagnato da percentuali relativamente basse di silicati; sono anche presenti alcune sostanze di identificazione meno chiara, ma il composto reattivo principale è l'alluminato monocalcico.

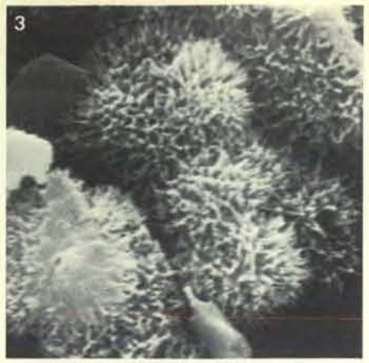
In entrambi i cementi i principali componenti anidri presentano proprietà idrauliche: finemente macinati essi reagiscono con l'acqua prima solidificando e poi dando luogo a una massa dura. Inoltre, poiché i cementi hanno una composizione chimica differente, il modo e la velocità con cui fanno presa e sviluppano la loro resistenza sono diversi. Entrambi i tipi possono raggiungere la stessa resistenza finale, ma il cemento alluminoso inizia a fare presa molto più rapidamente del cemento Portland. In pratica il primo può raggiungere in un giorno all'incirca la stessa resistenza raggiunta in un mese da un analogo cemento del secondo tipo. Sotto questo punto di vista i vantaggi del cemento alluminoso sono evidenti; il suo svantaggio è però rappresentato dalla resistenza a lungo termine. Esso può degradarsi secondo un processo noto come conversione, dipendente dalla temperatura e dalle condizioni ambientali e per effetto del quale la resistenza meccanica manifesta un'improvvisa e impressionante diminuzione. Le condizioni che esaltano tale comportamento sono un'elevata temperatura ambiente e un troppo alto contenuto di acqua nell'impasto. Anche nel cemento Portland questi fattori tendono a ridurre la resistenza per qualsiasi stagionatura,



La microfotografia al microscopio elettronico a scansione con un ingrandimento di 6000 volte mostra la struttura del cemento Portland nello stadio iniziale della reazione con l'acqua. Dai granelli di cemento si sviluppa il prodotto della reazione costituito da fibre di gel di silicato.

Il cemento alluminoso, progettato per resistere all'attacco da parte di solfati e che indurisce più rapidamente del cemento Portland, è mostrato con lo stesso ingrandimento della illustrazione in alto. I prodotti di reazione sono alluminati cristallini aventi forme piatte e allungate che si intrecciano. La instabilità di questi prodotti di reazione è all'origine del problema presentato dal cemento alluminoso, cioè l'impossibilità di prevederne la resistenza a lungo termine.





Nelle tre microfotografie al microscopio elettronico a scansione sono mostrate malte di cemento Portland in vari stadi di idratazione, cioè in tempi diversi dopo l'impasto con l'acqua. Dopo due ore (*1*) intorno ai granelli di cemento sono visibili i rivestimenti iniziali di gel. Dopo un mese (*2*) sono evidenti sia le fibrille sia i cristalli piatti di idrato di calcio. Il dettaglio dello stesso campione (*3*) mostra le fibre intrecciate.

ma si tratta di effetti molto meno marcati che si verificano senza l'improvvisa caduta di resistenza che si ha nel cemento alluminoso.

Supponendo che la composizione del cemento Portland sia quella prima indicata, è opportuno prendere in considerazione le conoscenze acquisite sui prodotti di idratazione e sulle loro proporzioni nel conglomerato cementizio. Come si può facilmente intuire, la reazione di presa in una malta di cemento e acqua è essenzialmente connessa con l'idratazione dei due silicati, C_3S e C_2S. Questi costituiscono circa il 75 per cento in peso del prodotto cotto, o clinker anidro, e forniscono la maggior parte del materiale idratato. Le velocità di reazione del C_3S e del C_2S differiscono in modo sensibile, ma il prodotto finale è da considerare in entrambi i casi lo stesso: un gel colloidale di idrosilicato di calcio avente composizione e struttura non precisata. In un cemento indurito questo gel, sotto forma di massa finemente suddivisa e prevalentemente amorfa, occupa circa il 70 per cento in volume del materiale idratato e costituisce quindi il principale agente legante fra i granelli di cemento residui che non hanno partecipato alla reazione e gli altri prodotti cristallini di idratazione. La maggior parte degli altri prodotti cristallini è costituita da idrato di calcio (portlandite), mentre la rimanente frazione è composta da un certo numero di idrati complessi di alluminio e da altre sostanze di minore importanza.

Dei due silicati il C_3S è soprattutto responsabile della reazione di indurimento che si verifica durante gli stadi iniziali dell'idratazione, entro i primi giorni e settimane. Al C_2S, con un maggior tempo di reazione, è probabilmente da attribuire lo sviluppo continuo e a lungo termine della resistenza meccanica, che può prolungarsi per mesi e perfino per anni. Il contributo dei composti alluminati alla resistenza non è accertato, ma è verosimilmente alquanto secondario.

In una malta di cemento la reazione di idratazione non perviene, in condizioni normali, al completamento. Perfino dopo parecchi anni una microstruttura tipica presenta granelli residui di cemento che non ha partecipato alla reazione, incastrati in una matrice composta in prevalenza da un gel di idrosilicato di calcio. La ragione di questo fatto si trova nel rivestimento che i prodotti di reazione formano intorno ai granelli di cemento durante l'idratazione, rivestimento che impedisce progressivamente il contatto dell'acqua con il materiale anidro. Solo nel caso di una continua soppressione di questi rivestimenti mediante una incessante agitazione della malta in un mulino a pale, per esporre superfici reattive fresche, si può ottenere che l'idratazione venga completata in un ragionevole lasso di tempo.

La presenza di clinker che non ha reagito può avere alcuni effetti interessanti. Per esempio il cemento indurito, se viene macinato e mescolato di nuovo con acqua, fa presa una seconda volta e raggiunge un grado di resistenza meccanica che, sebbene inferiore a quello della prima presa, può essere non di meno notevole. È anche possibile che in una struttura di cemento si sviluppino delle sottili microfessurazioni, eliminabili con una nuova idratazione.

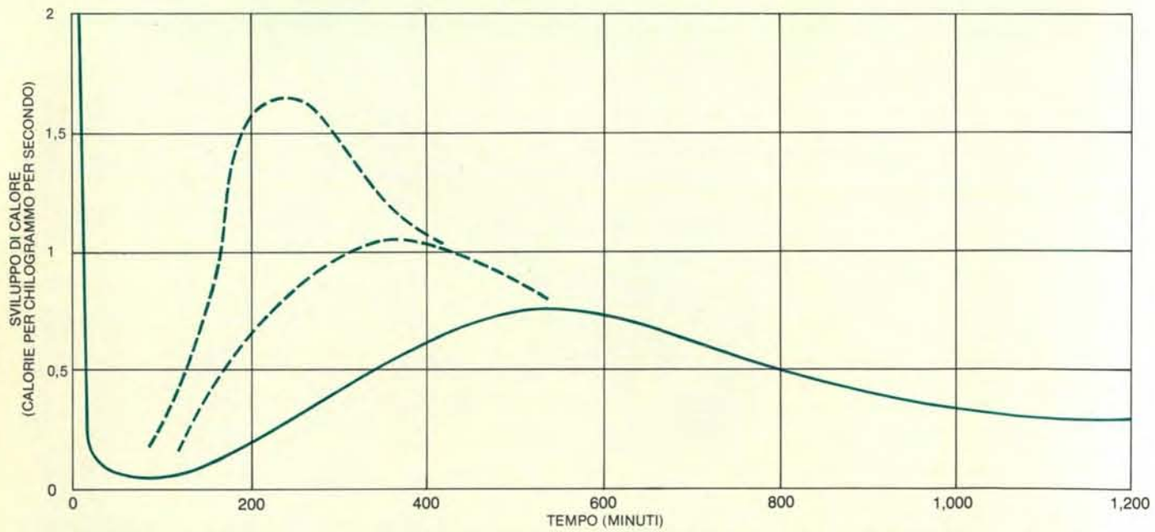
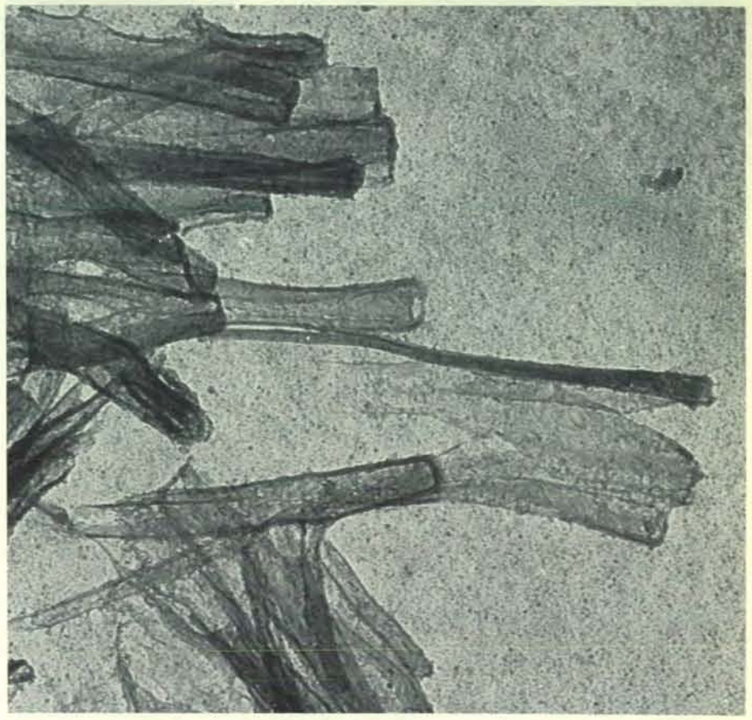


Diagramma dello sviluppo di calore durante la presa e l'indurimento del cemento Portland. Il primo massimo molto pronunciato è dovuto ai rapidi processi esotermici iniziali, compresa l'idratazione della calce libera e lo scioglimento delle impurità presenti nel cemento. Il minimo successivo corrisponde al periodo «latente» o di «induzione» dovuto probabilmente al rallentamento dell'idratazione provocato dai rivestimenti iniziali di gel che si formano intorno ai granelli di cemento. In un tempo successivo i rivestimenti cessano la loro azione perché la velocità di reazione comincia di nuovo a salire per raggiungere ben presto un secondo massimo. A una temperatura di idratazione di 20 gradi centigradi la velocità di reazione è rappresentata dalla curva inferiore. Le curve tratteggiate superiori mostrano l'effetto che un aumento della temperatura di idratazione, rispettivamente a 30 e a 40 gradi centigradi, ha nei confronti della forma del secondo massimo.

La reazione fra il cemento e l'acqua è esotermica; in condizioni tipiche la temperatura della malta può aumentare di decine di gradi. È un fenomeno che può essere un difficile problema durante le gettate per strutture di grande sezione. Talvolta si rende necessario raffreddare il materiale durante i periodi di presa e di indurimento per evitare fessurazioni dovute alla dilatazione termica e al successivo ritiro. Il calore sviluppato è direttamente proporzionale al grado di idratazione dei vari componenti del cemento. A causa della natura fisica della reazione, la velocità di emanazione del calore si riduce esponenzialmente a lunghi intervalli di tempo. Tuttavia in condizioni normali il 50 per cento circa del calore disponibile complessivo (da 100 a 150 calorie per grammo) viene liberato entro i primi tre giorni dall'impasto; nel primo periodo ne viene liberato grosso modo il 30 per cento.

La curva dello sviluppo del calore (*si veda l'illustrazione in basso nella pagina a fronte*) presenta due massimi. Il primo è netto e corrisponde ai processi iniziali fortemente esotermici; esso viene raggiunto entro 10 minuti dall'impasto ed è dovuto in modo prevalente all'idratazione della calce libera e allo scioglimento di altre impurità sempre presenti in piccole quantità nel cemento Portland commerciale. Raggiunto il primo massimo, la velocità di reazione scende rapidamente a un valore inferiore e vi si mantiene per un'ora o due: il corrispondente periodo è detto di «induzione» o «latente». Si pensa che tale fenomeno sia dovuto al rallentamento dell'idratazione provocato dal rivestimento dei granelli di cemento da parte dei prodotti gelatinosi iniziali; l'ipotesi è confermata dalle osservazioni al microscopio elettronico.

In un tempo successivo i rivestimenti cessano apparentemente la loro azione



Microstruttura del cemento Portland idratato vista al microscopio elettronico a trasmissione. La fotografia in alto mostra un campione umido di cemento dopo circa due giorni di idratazione; si osservano con un ingrandimento di 20 000 volte gli accrescimenti fibrillari ben sviluppati del gel di silicato intorno ai granelli di cemento. Quella in basso, con un ingrandimento di 70 000 volte, mostra un campione essiccato di cemento idratato nel quale si distinguono le fibre tubolari.

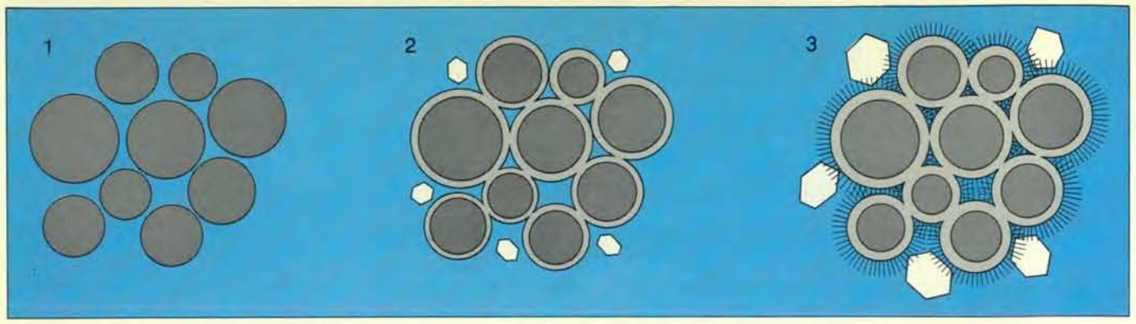


Illustrazione schematica degli stadi di idratazione del cemento Portland. Il processo comincia (*1*) con l'aggiunta di acqua che circonda i singoli granelli di cemento. Ben presto intorno ai granelli compare (*2*) il rivestimento di gel e si presentano i cristalli angolosi di idrato di calcio che si sviluppa come prodotto secondario dell'idratazione dei silicati presenti nell'impasto. Infine si sviluppano (*3*) le fibrille. Con il procedere dell'idratazione l'interconnessione di queste fibrille lega fra loro il cemento e gli altri componenti per formare una massa dura.

perché la velocità di reazione comincia di nuovo a salire per raggiungere un secondo massimo (di valore molto minore) in un periodo compreso fra otto e dieci ore, dopo di che decresce con gradualità in un intervallo molto più lungo. La forma del massimo è in gran parte determinata dall'idratazione del componente principale del cemento, il silicato tricalcico. La curva ha l'andamento caratteristico mostrato da molte fasi di trasformazione che si verificano mediante un processo di germinazione e di accrescimento. Come è logico, il grado di reazione aumenta con la temperatura ed è dipendente direttamente dalla superficie della polvere di cemento, cioè dalla finezza con cui è stato macinato il clinker. Le variazioni nel rapporto fra acqua e cemento entro limiti compresi fra il 30 e il 50 per cento (come si verifica nella pratica corrente) non alterano in modo sensibile l'andamento dell'idratazione.

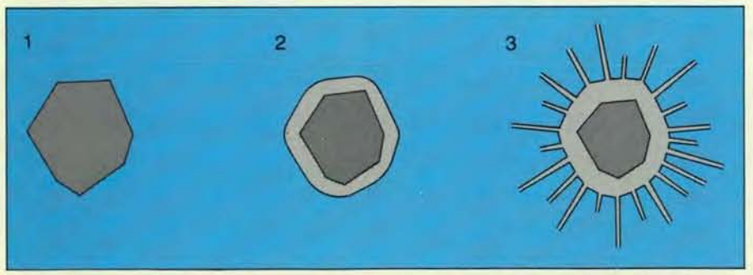
Sebbene le curve dello sviluppo di calore forniscano alcune informazioni sul processo di idratazione, dicono ben poco sulla struttura del materiale e non sono necessariamente connesse in modo semplice all'aumento della resistenza meccanica. È vero che in condizioni normali il periodo di presa corrisponde pressappoco all'intervallo di tempo fra lo stadio latente e il secondo massimo e che a esso segue la reazione di indurimento, ma i processi di presa e di indurimento dipendono con evidenza da come si sviluppa nei particolari la microstruttura della malta. La microstruttura degli idrati interessati è stata studiata con il migliore profitto, a causa della loro fine granulometria, mediante la microscopia elettronica.

Le microfotografie mostrano che entro due ore dall'impasto i granelli cristallini e angolati del cemento vengono rivestiti e collegati fra loro da involucri gelatinosi di materiale idrato. Il reticolato di legami relativamente deboli che così si forma spiega la perdita di plasticità della malta in questa fase, dovuta cioè all'effetto della presa. La struttura può essere distrutta con facilità, ma si riforma ben presto con una nuova coagulazione dei granelli di cemento.

Lo sviluppo irreversibile della resistenza meccanica durante il periodo di indurimento, che inizia da tre a cinque ore dopo l'impasto, è connesso con un netto cambiamento nella microstruttura del cemento. Dai rivestimenti iniziali di gel si sviluppano delle sottili protuberanze superficiali che in un certo intervallo di tempo si accrescono assumendo la forma di «fibrille» fittamente distribuite che si irradiano dai singoli granelli di cemento in modo simile agli aculei di un istrice. Tale sviluppo fibrillare è il prodotto della reazione fra i silicati del cemento e l'acqua, cioè il gel di idrosilicato di calcio. Nel contempo l'idrato di calcio (portlandite) precipita come sottoprodotto e forma dei grossi cristalli angolosi, di grande rilievo nella microstruttura. Con il procedere dell'idratazione le fibrille pervadono progressivamente la regione fra i granelli adiacenti di cemento. La maglia intrecciata che così si forma alla fine si consolida formando una matrice rigida che tiene unita con efficacia la massa composita costituita dai granelli residui di cemento che non hanno partecipato alla reazione, dai prodotti cristallini di idratazione e, in un calcestruzzo, dalla sabbia e dal materiale inerte.

Malgrado l'evidente importanza che la crescita secondaria delle fibrille di gel di idrosilicato di calcio ha per lo sviluppo della resistenza meccanica nell'impasto cemento e acqua, la morfologia dettagliata e il meccanismo del fenomeno sono rimasti finora oscuri. Recenti studi eseguiti a Oxford dal gruppo di ricerca degli autori con l'uso del microscopio elettronico a trasmissione hanno apportato qualche nuova conoscenza sul problema. Le immagini ottenute dimostrano che le fibrille non hanno le forme regolari e sfaccettate che ci si potrebbero attendere dall'accrescimento di cristalli solidi da una soluzione acquosa e non sono piene ma sembrano piuttosto essere costituite da sottili strutture cave tubolari. Sono informazioni che danno al problema un aspetto del tutto nuovo.

Nei fenomeni di precipitazione dei solidi da una soluzione acquosa il concetto di una morfologia tubolare è forse alquanto insolito. È risultato tuttavia che tali forme di accrescimento sono più comuni di quanto non ci si aspettasse. Un esempio familiare di accrescimento tubolare può essere trovato nei «giardini» silicei che si formano nel caso di



I «giardini» silicei offrono un'analogia con l'accrescimento delle fibrille tubolari nel cemento Portland. Un giardino siliceo si forma quando sali solubili in acqua di alcuni metalli vengono disciolti in una soluzione di silicato di sodio. Nell'esempio un cristallo di nitrato di cobalto, comincia a sciogliersi (*1*) e intorno a esso si forma un involucro di silicato insolubile (*2*). Poiché la membrana è permeabile all'acqua, il sale continua a passare in soluzione. Infine la pressione osmotica apre la membrana (*3*), provocando l'accrescimento del precipitato di silicato in forma di sottili tubi cavi analoghi a quelli del cemento Portland idratato, anche se di dimensioni minori.

Sviluppo della microstruttura in una malta di cemento alluminoso, mostrato al microscopio elettronico a scansione. La malta è stata realizzata con un elevato rapporto (60 per cento) fra acqua e cemento e fatta stagionare ad alta temperatura (40 gradi centigradi) per studiare il processo di conversione che spesso determina il cedimento sotto carico di elementi realizzati in cemento alluminoso. In alto si vede la malta dopo un giorno dall'impasto; sono evidenti i cristalli aghiformi dei prodotti iniziali metastabili di idratazione. In basso, dopo 14 giorni di idratazione, si vedono i poliedri sfaccettati del prodotto di conversione stabile, ma debole.

soluzioni acquose diluite di silicato di sodio in cui vengono versati sali solubili in acqua di metalli che a loro volta formino silicati insolubili. La configurazione di accrescimento di un giardino siliceo offre una notevole analogia visiva con quanto avviene durante l'idratazione del cemento Portland. Intorno al cristallo di sale precipita all'inizio un involucro gelatinoso di idrosilicato. Dopo un periodo latente nell'involucro di gel si formano delle aperture da cui si protendono delle sottili formazioni tubolari di accrescimento che hanno l'aspetto di fibrille radiali. Malgrado alcune ovvie differenze fra l'idratazione del cemento Portland e l'accrescimento di un giardino siliceo (nella velocità di accrescimento e nella natura chimica dei componenti) le analogie fra i due processi sono tanto strette da suggerire che in entrambi i casi il procedimento di crescita sia dello stesso tipo.

È stato dimostrato che il meccanismo di accrescimento del giardino siliceo dipende dalla pressione osmotica. I cristalli di sale cominciano a sciogliersi quando vengono in contatto con la soluzione acquosa di silicato, mentre nello stesso tempo fanno precipitare un involucro gelatinoso di idrosilicato. L'involucro è colloidale e si comporta come una membrana semipermeabile: per la differenza di concentrazione fra le soluzioni delle due parti della membrana (una soluzione concentrata di sale metallico all'interno e una diluita di silicato all'esterno) l'acqua tende a passare all'interno della membrana, determinando un continuo passaggio in soluzione dei cristalli di sale. La pressione interna aumenta gradualmente finché determina l'apertura dell'involucro con la eiezione di getti di soluzione salina nella soluzione di silicato. Fra le due soluzioni avviene una reazione spontanea con la precipitazione del silicato insolubile sotto forma di una parete tubolare in continuo accrescimento. La membrana iniziale agisce come una pompa osmotica nella quale la soluzione salina si sposta lungo il tubo e precipita alla sua estremità in un processo quasi stazionario.

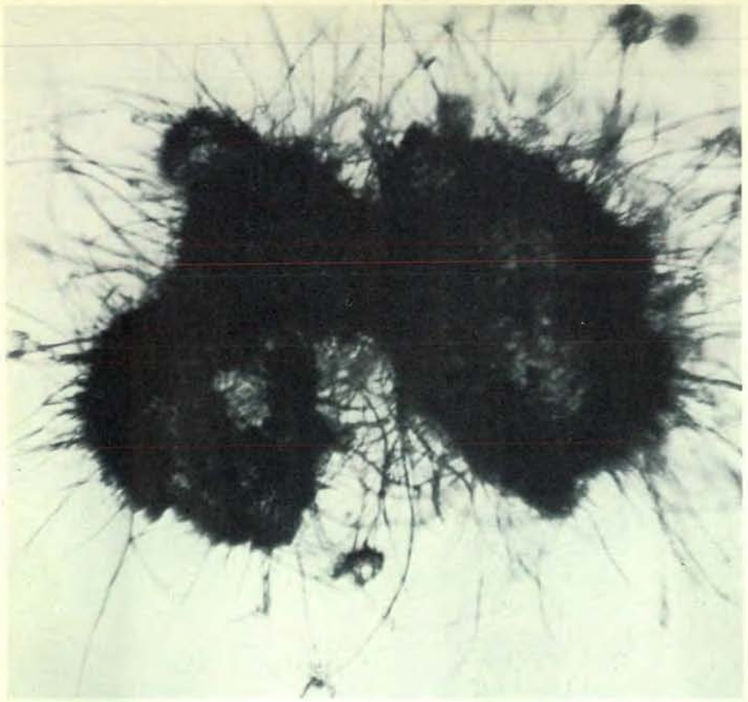
Su tale modo di accrescimento sono finora disponibili pochi dati quantitativi, ma è probabile che esso sia relativamente frequente ogni qualvolta il prodotto iniziale della reazione fra due soluzioni sia una membrana continua semipermeabile. Sia l'involucro iniziale sia i tubi che da esso si accrescono sono amorfi e gelatinosi, malgrado che si dimostrino abbastanza duri e resistenti quando vengono estratti dalla soluzione. In ogni sistema la velocità di accrescimento è proporzionale al diametro del tubo; i tubi di minori dimensioni sono il risultato di accrescimenti più lenti. L'ultima osservazione conferma dal punto di vista qualitativo l'analogia fra i giardini silicei e i prodotti di idratazione del cemento Portland. Un gran numero di sali di ioni metallici presentano questo tipo di accrescimento e fra essi è rassicurante trovare i sali di calcio, poiché il gel di idrosilicato di calcio è il principale prodotto di idratazione del cemento Portland.

Prendendo a modello il giardino siliceo, è possibile spiegare le caratteristiche essenziali dell'idratazione del cemento Portland e in particolare l'accrescimento del materiale fibrillare costituito da gel di idrosilicato di calcio, che è un fattore importantissimo nello sviluppo della resistenza meccanica. L'analogia implica che sia l'osmosi a fornire la forza inducente l'idratazione del cemento Portland. Sono in corso dettagliati studi dai quali si spera di ottenere una conoscenza più approfondita sulla natura fondamentale di questo materiale.

I prodotti di idratazione del cemento alluminoso hanno forme cristalline più convenzionali. Come è stato indicato, il principale componente anidro del cemento alluminoso è l'alluminato monocalcico ($CaO \cdot Al_2O_3$, scritto anche CA) e lo sviluppo della resistenza meccanica deriva soprattutto dai prodotti di idratazione di questo composto. In condizioni normali e a temperatura ambiente, il prodotto principale è la fase decaidrata CAH_{10}, associata in generale con quantitativi minori di ottaidrato C_2AH_8 con gel di ossido di alluminio, il quale alla fine si trasforma durante la stagionatura in gibbsite, o idrargillite, cristallina, AH_3. Le quantità relative di CAH_{10} e di C_2AH_8 presenti nella malta di cemento dipendono dalla temperatura di idratazione e dalla composizione di partenza del cemento stesso. Gli idrati cristallini costituiscono la matrice che lega la malta e la loro rapida formazione dà luogo al veloce aumento della resistenza meccanica, caratteristico del materiale.

La conversione si verifica perché gli idrati iniziali CAH_{10} e C_2AH_8 sono metastabili e tendono a trasformarsi spontaneamente nell'idrato C_3AH_6 più stabile, con liberazione di acqua e di altra gibbsite quali prodotti secondari di reazione. Il processo è accompagnato da un cambiamento nella struttura cristallina dell'idrato (CAH_{10} e C_2AH_8 hanno strutture cristalline esagonali ben definite, mentre C_3AH_6 ha struttura cubica) e da una sensibile riduzione di volume per la differenza di densità dei solidi. La riduzione arriva quasi al 50 per cento per il CAH_{10} e al 30 per cento circa per il C_2AH_8.

Questi cambiamenti determinano il problema della resistenza a lungo termine del cemento alluminoso in quanto conducono a una perdita di coesione della microstruttura e a un aumento di porosità della massa di cemento indurita. A temperatura uguale o inferiore a quella ambiente la velocità delle variazioni è bassa, soprattutto in climi secchi, e gli effetti possono non manifestarsi per anni. A temperature più elevate e in ambienti umidi la conversione può verificarsi nel giro di poche settimane o perfino di alcuni giorni. Il risultato finale è una marcata riduzione della resistenza meccanica e una maggiore soggezione all'attacco di sostanze corrosive. Il cemento alluminoso dura più a lungo se all'atto dell'impianto il clinker è molto secco e la temperatura è bassa. In condizioni contrarie non si può ottenere la



L'intreccio di sottili fibrille tubolari suggerisce una possibile struttura migliorata del cemento Portland. Il cristallo di solfato ferroso, ingrandito 32 volte, è stato immerso in una soluzione di silicato di sodio. Un più fitto intreccio delle fibre migliorerebbe la resistenza a trazione.

stabilità a lungo termine del materiale.

È possibile seguire le reazioni esaminando al microscopio elettronico a scansione campioni di malta di cemento in vari stadi di idratazione. I campioni esaminati dagli autori, ottenuti con un elevato rapporto fra acqua e cemento, erano stati stagionati artificialmente in un ambiente umido mantenuto a temperatura abbastanza alta (40 gradi centigradi). All'inizio il cemento viene ricoperto da uno strato liscio e gelatinoso, di aspetto non molto diverso da quello che si forma sui granelli di cemento Portland. Dopo un breve periodo di incubazione gli involucri vengono aperti da creste che si accrescono rapidamente sotto forma di cristalli sfaccettati con contorni esagonali o aghiformi. Questi cristalli possono essere identificati con i primi prodotti di idratazione: grazie alla loro complessa interconnessione il cemento sviluppa il rapido aumento di resistenza.

Nelle condizioni di stagionatura escogitate, il campione subisce una rapida trasformazione ed entro 14 giorni dall'impasto presenta il prodotto di idratazione più stabile: l'esaidrato in forma quasi sferica con poliedri sfaccettati. Poiché i cristalli sfaccettati sono caratteristici degli accrescimenti da soluzioni, si conclude che la conversione in idrato stabile deve avvenire per questa via. In altre parole il cambiamento implica una ricristallizzazione da una soluzione acquosa. È quindi evidente nella microstruttura del cemento la presenza di acqua necessaria per fornire il mezzo nel quale la fase stabile può germinare e accrescersi. Questa conclusione suggerisce una spiegazione al fatto che un elevato rapporto fra acqua e cemento può facilitare la perdita di resistenza nel cemento alluminoso.

Si è portati a chiedersi che cosa si può fare a proposito di questo problema. È possibile impedire la trasformazione oppure indurre la formazione immediata del prodotto più stabile? È difficile immaginare come possa essere impedita la trasformazione salvo mediante la pratica consueta di tenere l'impasto piuttosto secco, in modo che la quantità d'acqua disponibile per provocare l'indesiderata trasformazione sia ridotta al minimo. Forse se l'idratazione iniziale venisse eseguita a temperatura sufficientemente alta, l'idrato stabile potrebbe svilupparsi con maggiore velocità. Però nella pratica le temperature necessarie sarebbero con molta probabilità troppo alte e in ogni caso in una matrice di cemento costituita da poliedri sfaccettati sarebbe ben difficile la formazione di quel reticolato di compenetrazione che sembra essere un requisito essenziale per un'elevata resistenza meccanica.

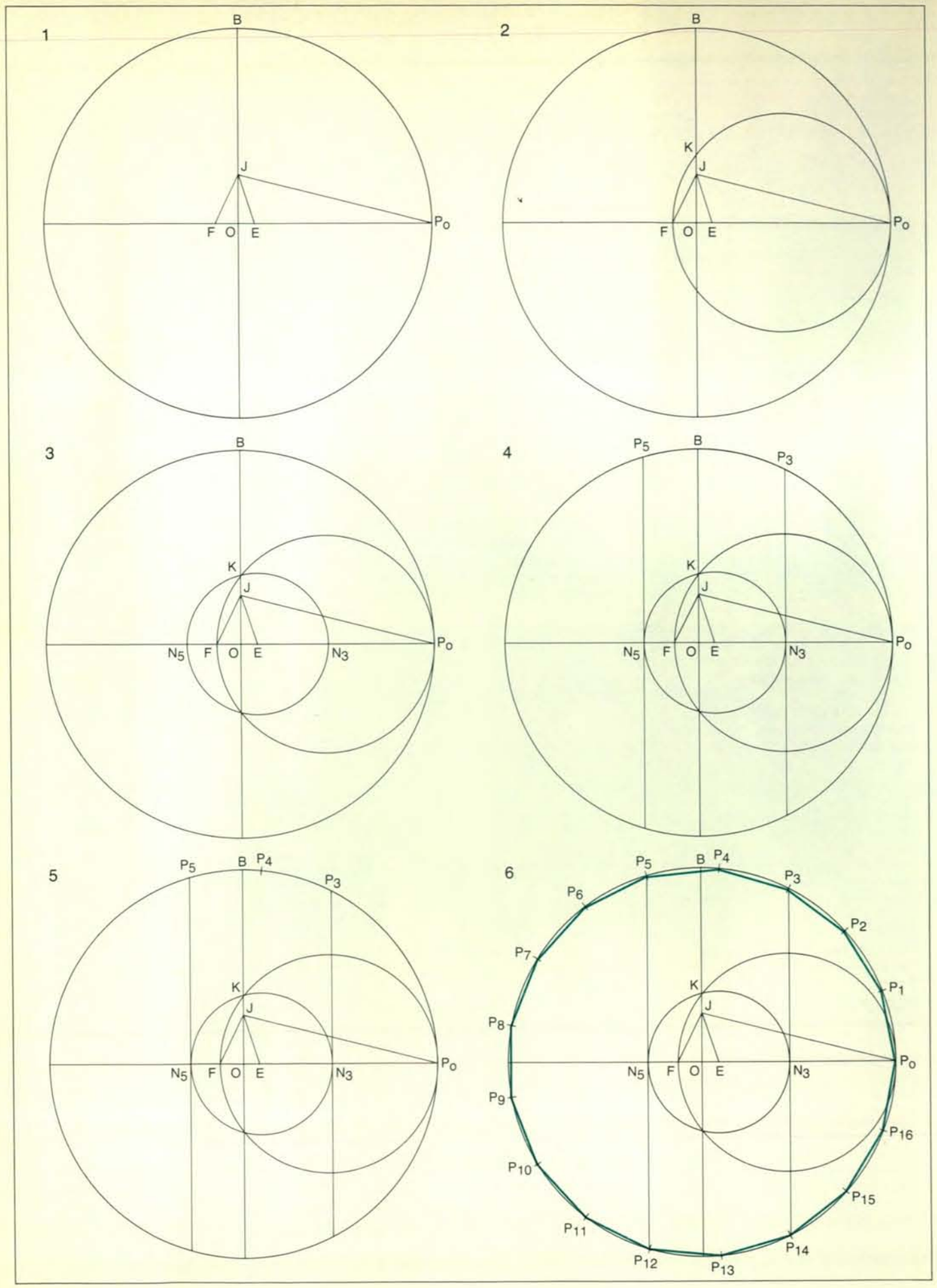
Altri problemi pratici si presentano poi negli elementi strutturali, quali una trave di cemento armato precompresso. Sebbene sia riconosciuto che il miglior modo di affrontare il problema sia di ridurre al minimo la quantità di acqua durante l'impasto del cemento, quale uniformità può raggiungersi in un elemento lungo da 5 a 10 metri e di sezione quadrata con lato da 25 a 30 centimetri? Può essere istruttivo prendere in esame un esempio.

Da una trave in cemento alluminoso lunga sei metri è stata prelevata una sezione dello spessore di due centimetri. Nel fondo della trave era inserito un elemento in legno, presumibilmente per inchiodarvi degli oggetti. Il campione è stato quindi tagliato in spezzoni di minori dimensioni, per ciascuno dei quali è stata determinata con analisi termochimica la relativa conversione (dal composto metastabile a quello stabile).

Come prevedibile in una struttura eterogenea del genere la distribuzione era variabile con alcune zone di bassa conversione in prossimità delle superfici e intorno all'elemento in legno, dove il calcestruzzo era stato localmente essiccato per l'assorbimento dell'acqua. Si deve inevitabilmente supporre che le modifiche nell'altra dimensione (secondo la lunghezza della trave) presentino una maggiore variabilità. Dall'esempio è possibile valutare i problemi pratici posti dalla necessità di assicurare l'omogeneità di un materiale del genere entro limiti precisi e nelle strutture di grandi dimensioni. Le variazioni nei valori di conversione devono riflettersi anche in variazioni della resistenza meccanica nel calcestruzzo da punto a punto. È da ricordare che in definitiva la resistenza di una trave dipende dal suo punto più debole.

Una delle proprietà caratteristiche dei cementi (e dei calcestruzzi in generale) è la forte resistenza alla compressione durante l'idratazione, mentre la loro resistenza alla trazione resta relativamente bassissima. Per questa ragione è di solito necessario armare le travi in calcestruzzo e sottoporre le armature a una sollecitazione di tensione opportunamente applicata prima o durante il disarmo della struttura. È chiaro che qualsiasi miglioramento ottenibile nella resistenza alla trazione di un cemento o di un calcestruzzo avrebbe un'importanza essenziale nell'ingegneria civile.

Se i processi di presa e di indurimento dipendono dalla compenetrazione e dal concatenamento di fibrille gelatinose e di cristalli sfaccettati, come suggeriscono le recenti scoperte prima descritte, è ragionevole chiedersi se la resistenza alla trazione possa essere migliorata esaltando in qualche modo lo sviluppo delle caratteristiche di compenetrazione, forse provocando l'accrescimento di prodotti che consentano un più stretto collegamento o una connessione più intima. Le prospettive di utili modifiche sono incerte con gli idrati cristallini ben sviluppati del cemento alluminoso. È invece possibile attendersi qualche progresso con le fibre irregolari che si riscontrano nel cemento Portland se, per esempio, queste potessero essere rese meno uniformi e più strettamente interconnesse.



La costruzione geometrica di un poligono regolare di 17 lati usando riga e compasso avviò Gauss alla sua carriera di matematico nel 1796, all'età di 18 anni. Una versione semplificata di questa costruzione, progettata da H.W. Richmond nel 1893, è riportata nella pagina a fronte. Le istruzioni sono le seguenti: (1) si tracci una circonferenza di centro *O* e raggio OP_0 di lunghezza arbitraria. Si tracci *OB* perpendicolare a *OP*. Si trovi il punto *J* a un quarto della distanza *OB*. Si trovi il punto *E* tale che l'angolo *OJE* sia un quarto dell'angolo OJP_0. (Ciò si può ottenere bisecando due volte l'angolo.) Si trovi il punto *F* tale che l'angolo *FJE* sia di 45 gradi. (Ciò si ottiene bisecando un angolo retto.) (2) Si tracci una circonferenza di diametro FP_0; questo cerchio interseca *OB* nel punto *K*. (3) Si tracci un'altra circonferenza di centro *E* e raggio *EK*. Tale circonferenza individua i punti N_5 ed N_3. (4) Si traccino i segmenti N_3P_3 e N_5P_5 perpendicolari a OP_0. (5) Si bisechi l'arco P_3P_5 per ricavare il punto P_4. (6) Usando la corda P_4P_5, si riporti successivamente tale lunghezza partendo da P_0. Si colleghino i punti ottenuti.

Gauss

Bambino prodigio che divenne il principale matematico del suo tempo, si trovava a suo agio sia con le astrazioni della teoria dei numeri sia con i lunghi calcoli dell'astronomia e le applicazioni pratiche della fisica

di Ian Stewart

«La matematica è la regina delle scienze», disse un giorno Carl Friedrich Gauss, e la sua carriera fu una dimostrazione della verità di tale aforisma. Considerato generalmente, insieme con Archimede e Newton, uno dei più abili matematici di tutti i tempi, Gauss si interessava sia della teoria che delle applicazioni e i suoi contributi vanno dalla teoria dei numeri a problemi pratici di astronomia, di magnetismo e di geodesia. In tutti i rami della matematica dei quali si occupò fece importanti scoperte, introdusse idee e metodi nuovi e gettò le basi per successive ricerche. A conferma delle sue grandi capacità, in quest'anno, il duecentesimo anniversario della sua nascita, molte delle sue idee stanno ancora dando frutti.

Gauss era sotto molti aspetti una personalità enigmatica e contraddittoria. Figlio unico appartenente a una famiglia di operai, egli riuscì a diventare il principale matematico del suo tempo, ma viveva modestamente ed evitava la pubblicità. Il suo carattere era mite, pur essendo nello stesso tempo un uomo distaccato, politicamente conservatore e spesso ostinato, che chiedeva soltanto di poter continuare indisturbato il suo lavoro creativo. Egli era disposto a riconoscere le capacità matematiche dovunque le trovasse, nonostante i pregiudizi dell'epoca, ma l'aver casualmente ignorato alcuni dei migliori giovani matematici del suo tempo, in particolare János Bolyai, uno dei pionieri della geometria non euclidea, ebbe conseguenze sfortunate.

Un aspetto particolarmente indicativo del carattere di Gauss era il rifiutarsi di presentare qualsiasi lavoro finché non era sicuro di averlo limato alla perfezione. Nessun risultato, per quanto importante, fu pubblicato fin quando non lo riteneva completo. Egli rielaborò le sue dimostrazioni matematiche a un livello tale che i passaggi attraverso i quali aveva ottenuto i suoi risultati erano quasi invisibili. I suoi lavori pubblicati hanno un tocco di grazia ed eleganza classiche che li rende austeri e ineguagliabili. Molte delle sue migliori idee non compaiono esplicitamente nei suoi scritti ma devono essere ricavate percorrendo a ritroso i passi attraverso i quali le sue scoperte dovevano esser state fatte. Il risultato fu che importanti concetti non videro la luce del giorno finché non vennero scoperti indipendentemente da altri.

Gauss pubblicò durante la sua vita circa 155 lavori, ma lasciò un gran numero di lavori non pubblicati. In questa breve rassegna accennerò alle sue più importanti scoperte e tenterò parzialmente di spiegare come ci sia arrivato.

Gauss nacque il 30 aprile del 1777 a Brunswick (ora nella Germania Occidentale). Il padre lavorava ora come giardiniere, ora come addetto ai canali, ora come muratore. Il figlio lo descrisse più tardi come «un uomo estremamente onesto e sotto molti aspetti stimabile e veramente rispettabile, ma in casa... ruvido e autoritario.» La madre di Gauss, figlia di uno scalpellino, era una donna intelligente dal carattere forte e suo fratello Friedrich ebbe una parte importante nella vita del giovane Gauss. Friedrich lavorava come tessitore di damaschi fini, ma i suoi interessi erano insolitamente vasti e dedicò molto tempo a incoraggiare Gauss e a raffinarne i metodi di ragionamento.

Tra i grandi matematici ve ne sono tanti che hanno mostrato il loro talento fin dall'infanzia e altrettanti che non ne hanno mostrato affatto fino alla maggiore età. Gauss fu indubbiamente il più precoce di tutti. Si divertiva a raccontare che sapeva contare ancor prima di parlare e molti aneddoti confermano queste sue doti eccezionali. Si dice che il giorno prima di compiere i tre anni suo padre stava preparando le buste paga settimanali per i lavoratori alle sue dipendenze, senza accorgersi che suo figlio lo stava osservando attentamente. Finiti i conteggi, il vecchio Gauss udì con stupore una vocina che diceva: «Papà, il calcolo è sbagliato, dovrebbe essere...» Controllati i conti, la cifra detta dal bambino risultò esatta. La cosa più notevole di questa storiella è che nessuno gli aveva insegnato l'aritmetica.

Altri aneddoti riguardano la costante precocità di Gauss. All'età di 10 anni, egli fu ammesso al corso di aritmetica. Il maestro propose un problema di questo tipo: qual è la somma di $1+2+3+\ldots+$ $+100$, dove ci sono 100 numeri e la differenza tra ogni numero e il successivo è sempre uno. C'è un semplice trucco per calcolare tale somma, che era noto al maestro, ma non ai suoi scolari.

Di norma il primo ragazzo che risolveva un problema metteva la propria lavagna con la risposta sulla cattedra del maestro, il ragazzo successivo metteva la propria lavagna su quella precedente e così via. Il maestro aveva appena finito di dettare il problema quando Gauss mise la sua lavagna sulla cattedra. «Fatto,» egli disse. Nell'ora successiva restò seduto con le braccia conserte, ricevendo di tanto in tanto una occhiata scettica da parte del maestro, mentre gli altri ragazzi eseguivano l'addizione. Alla fine dell'ora il maestro esaminò le lavagne. Su quella di Gauss c'era un solo numero. Anche da vecchio a Gauss piaceva rac-

contare che di tutte le risposte la sua era l'unica a essere esatta.

A onore del maestro va detto che egli restò talmente impressionato da comprare con i propri soldi un libro di aritmetica per donarlo a Gauss che lo divorò rapidamente. Gauss ebbe anche la fortuna di avere un assistente di matematica del maestro, il diciassettenne Johann Martin Bartels, appassionato di matematica, così i due passarono molte ore a studiare insieme. Conosciuto il teorema binomiale, secondo il quale per qualsiasi numero n l'espressione $(1+x)^n$ è una serie e, se n non è un intero positivo, la serie è divergente, Gauss era insoddisfatto per la mancanza di rigore del libro che il maestro gli aveva donato e ne diede una dimostrazione. Pur essendo ancora uno scolaro, egli fu il primo matematico a prestare seria attenzione al problema degli infiniti. Pochi geni matematici hanno un'abilità che va al di là di una predisposizione nell'eseguire i calcoli, ma le capacità di Gauss si estendevano chiaramente a campi più vasti della conoscenza.

A 14 anni Gauss fu presentato al Duca di Brunswick, che aveva sentito parlare della sua fama e ne divenne il protettore. L'anno seguente Gauss entrò nel Collegium Carolinum di Brunswick, dove studiò le opere di Newton, di Leonhard Euler e di Joseph Louis Lagrange, assimilandole prontamente. Prima dei 19 anni aveva scoperto da solo e dimostrato un importante teorema della teoria dei numeri noto come legge di reciprocità quadratica (sulla quale tornerò più avanti). Per capire l'importanza di questo teorema si tenga presente che, pur avendone Euler intuito la validità in precedenza, sia lui che Adrien Marie Legendre non erano riusciti a dimostrarlo.

Quando Gauss lasciò il Collegium Carolinum nell'ottobre 1795 per studiare all'Università di Göttingen, era diviso tra la matematica e l'altra sua grande passione: lo studio delle lingue antiche, nelle quali era altrettanto brillante. Il 30 marzo del 1796 prese però una decisione in seguito a una delle più sorprendenti scoperte della storia della matematica.

Per avere qualche conoscenza di base su questa scoperta, torniamo indietro di duemila anni alla Grecia classica. Il principale contributo greco alla matematica veniva dalla fiorente scuola di geometria associata ai nomi di Pitagora, Eudosso, Euclide, Apollonio e Archimede. I greci furono probabilmente i primi a dare importanza al rigore nelle dimostrazioni e per mantenere tale rigore avevano imposto un certo numero di restrizioni. Una di esse era che nelle costruzioni geometriche si dovesse usare solo riga e compasso. In pratica le sole curve consentite erano la linea retta e il cerchio.

Euclide aveva dimostrato che con riga e compasso era possibile costruire poligoni regolari con tre, quattro, cinque e quindici lati, oltre ai poligoni ricavati con successive bisezioni di tali lati. Tali poligoni erano però i soli che i greci sapessero costruire; essi non conoscevano alcun metodo per costruire poligoni, per esempio, con 7, 9, 11, 13, 14 e 17 lati. Pare che per i successivi 2000 anni nessuno abbia sospettato che fosse possibile costruire qualcuno di tali altri poligoni. La scoperta di Gauss consisteva nella costruzione del poligono regolare di 17 lati iscritto nel cerchio usando soltanto riga e compasso. Inoltre egli spiegò esattamente quali altri poligoni si potevano costruire in tal modo: il numero dei lati doveva essere una qualsiasi potenza di 2 (2^n), oppure una potenza di 2 moltiplicata per uno o più differenti numeri primi dispari del tipo noto come primi di Fermat (da Pierre de Fermat, che li scoprì). Un numero primo è per definizione un numero maggiore di 1 che non può essere diviso esattamente per nessun numero tranne che per se stesso e per 1; un primo di Fermat ha l'ulteriore proprietà di essere maggiore di una unità rispetto a 2 elevato a una potenza di 2, cioè di tipo $2^{2^n}+1$. Gli unici primi di Fermat che si conoscano sono 3, 5, 17, 257 e 65 537. Abbiamo così l'importante risultato che i poligoni regolari con 17 lati si possono costruire con riga e compasso, mentre i poligoni regolari con 7, 9, 11, 13 e 14 lati non possono essere costruiti in tal modo.

Gauss dimostrò questo teorema (alla età di 18 anni) combinando un argomento algebrico con uno geometrico. Egli dimostrò che la costruzione di un poligono di 17 lati equivale alla soluzione dell'equazione $x^{16}+x^{15}+\ldots+x+1=0$. Dato che 17 è un numero primo e 16 è una potenza di 2 risulta che tale equazione può essere ridotta a una serie di equazioni quadratiche (espressioni della forma $ax^2+bx+c=0$, dove a, b e c sono numeri dati e x è l'incognita). Dal momento che era già stato dimostrato che le equazioni quadratiche si possono risolvere con una costruzione con riga e compasso, la dimostrazione era completa. A parte la importanza che tale dimostrazione ebbe nel convincere Gauss a dedicarsi interamente alla matematica, è il primo vero esempio (dopo l'introduzione delle coordinate a opera di Descartes) di una tecnica che doveva rivelarsi utilissima in matematica: spostare un problema da un campo (in questo caso la geometria) a un altro (l'algebra) e quivi risolverlo.

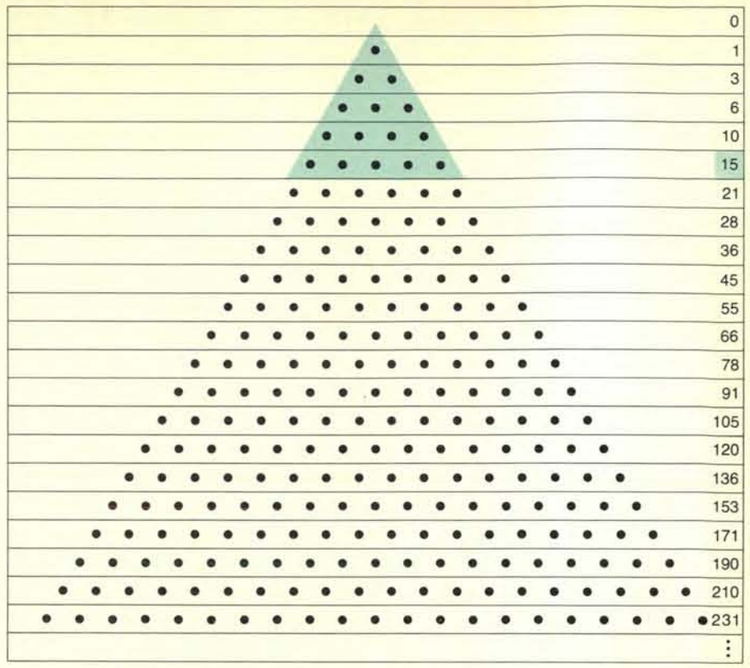
In questa elaborata incisione sono riprodotti i ritratti di Gauss (*in secondo piano*) e del fisico Wilhelm Weber (1804-91), con il quale collaborò in molti esperimenti di magnetismo e telegrafia. I motti greco e latino sul nastro suonano così: «Dio aritmetizza» e: «Il fine corona l'opera». La citazione greca più sotto, dall'inizio della *Repubblica*, dice: «Chi ha torce le passerà ad altri.»

Quando aveva vent'anni, scrisse un giorno Gauss, era talmente occupato da problemi matematici da non aver abbastanza tempo per prendere nota che di una piccola parte di essi. Molti di quelli che egli riuscì a risolvere apparvero nella sua opera *Disquisitiones Arithmeticae,* pubblicata nel 1801, all'età di 24 anni. Si può dire che questo lavoro rappresenta per la teoria dei numeri ciò che l'opera di Euclide rappresentava per la geometria: organizzava scoperte frammentarie sugli interi e le arricchiva con alcuni profondi concetti dello stesso Gauss. Gauss basò la sua teoria sul concetto di congruenza: due interi a e b sono detti congruenti «modulo» m se danno lo stesso resto quando vengono divisi per lo stesso numero m. Per esempio, 16 e 23 danno entrambi resto 2 quando vengono divisi per 7 e sono quindi congruenti modulo 7. Sette e nove sono congruenti modulo 2, dato che danno resto 1 quando entrambi vengono divisi per 2. Ovviamente presi comunque due numeri pari, o due numeri dispari, essi sono congruenti modulo 2.

Gauss indicò anche la possibilità di istituire un'aritmetica dei numeri congruenti. Egli dimostrò che se a, b, c e d sono interi e se dato come modulo un numero m, a è congruente a b e c è congruente a d, allora $a+c$ è congruente a $b+d$ e ac è congruente a bd. Possiamo quindi sostituire gli interi con i numeri congruenti nella nostra aritmetica senza incorrere in contraddizioni. Abbiamo però qualche sorpresa, come il fatto che, modulo 3, $1+1+1$ è congruente a zero.

Questa aritmetica dei numeri congruenti viene insegnata in molti corsi di «matematica moderna» con il nome di aritmetica modulare. Mi domando quanti insegnanti conoscono le sue origini e sanno perché Gauss la sviluppò. Egli se ne serviva come strumento per la dimostrazione di teoremi complessi e difficili. Forse il più importante di essi e senza dubbio quello preferito da Gauss (lo chiamava il teorema aureo) è la legge della reciprocità quadratica. A questo punto è necessario introdurre un po' di terminologia tecnica.

Gauss definì innanzitutto un «residuo quadratico» dicendo che se m è un intero positivo e a è un intero che non ha fattori in comune con m, allora a è un residuo quadratico di m, se è congruente, modulo m, a un quadrato perfetto. Un altro modo di dire la stessa cosa è il seguente: se a è un residuo quadratico di m, è possibile trovare almeno un numero x il cui quadrato, diviso per m, dia per resto a. Così 13 è un residuo quadratico di 17, perché l'affermazione «x^2 è congruente a 13 modulo 17» è soddisfatta se x assume il valore 8. Gauss dimostrò che se p e q sono numeri primi dispari differenti, allora p è un residuo quadratico di q se, e solo se, q è un residuo quadratico di p. C'è una sola eccezione a tale regola: se tanto p che q sono della forma $4n+3$, allora uno è un residuo quadratico e l'altro non lo è. Questo risultato può sembrare a prima vista estremamente specifico, ma ci permette di stabilire se un dato numero primo dispari è un residuo quadratico di un altro numero primo ponendo una domanda che in molti casi è più semplice: il secondo numero primo è un residuo quadratico del primo? Questo teorema ha ispirato molte profonde idee dell'algebra moderna ed è di grande importanza nella teoria dei numeri e in altri rami della matematica. Gauss lo ritenne talmente significativo da dimostrarlo in otto modi diversi.

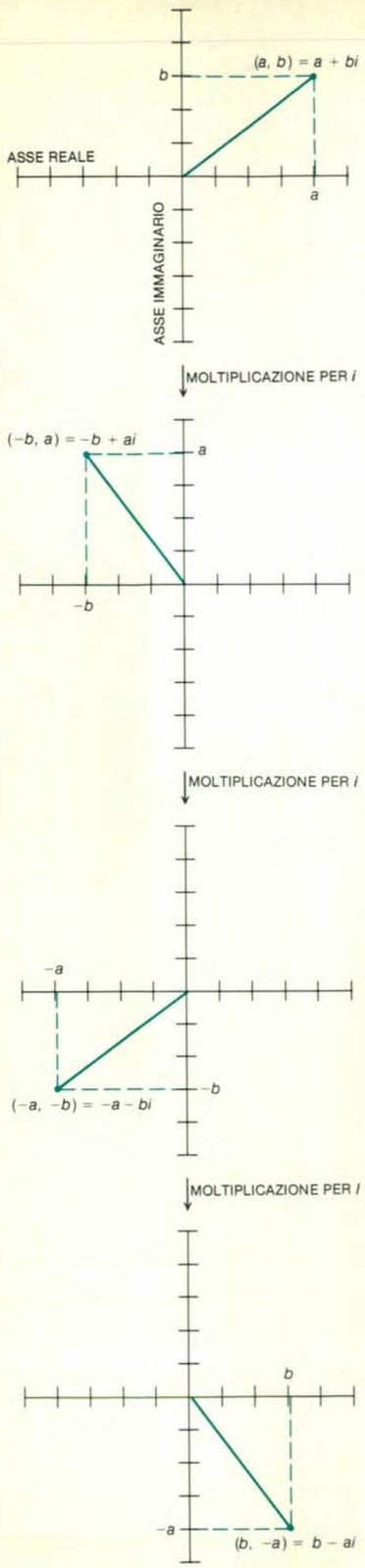


I numeri triangolari sono numeri della forma $n(n+1)/2$, dove n è un qualsiasi intero positivo, che si possono rappresentare con una disposizione triangolare di punti. Gauss dimostrò nelle sue *Disquisitiones Arithmeticae* che ogni intero positivo è la somma di tre numeri triangolari e annotò la scoperta nel diario il 10 luglio 1796 nel modo seguente: «Eureka! num = △ + △ + △.»

Le *Disquisitiones* misero in evidenza una tendenza che per Gauss diventò un modo di vivere. Le dimostrazioni venivano perfezionate al massimo grado, eliminando, se possibile, ogni traccia del processo con il quale erano state ottenute, in modo tale che rimaneva soltanto la forma finale. Gauss disse una volta: «Quando una bella costruzione è terminata, l'impalcatura non deve più vedersi.» Le generazioni successive, di fronte al problema di capire i metodi e i risultati di Gauss, possono essere perdonate se lo accusano di aver non soltanto tolto le impalcature, ma di aver anche gettato via i progetti. Nella ricerca matematica i concetti e i metodi sono spesso più importanti dei teoremi per i quali sono stati escogitati. Un'idea veramente buona può essere utilizzata in nuovi campi per fornire risultati che non si sarebbero potuti immaginare. A questo punto entrano in conflitto due aspetti della matematica: la matematica come forma artistica e la matematica come disciplina vivente.

Perché Gauss tenne nascosti i suoi metodi, preferendo dare soltanto una sintesi ed eliminando l'analisi? Egli adottò il motto *Pauca sed matura* che riflette la sua insoddisfazione per i teoremi incompleti dei suoi colleghi. Non possiamo esserne certi, ma forse l'indigenza durante la sua fanciullezza lo indusse a essere parsimonioso nella distribuzione delle sue idee. Può anche darsi che egli non volesse rendere pubblici lavori finiti a metà per timore di essere criticato se fossero risultati errati. Tale timore è comune ad altri grandi matematici; Newton, per esempio, dovette essere spinto a pubblicare i suoi *Principia.* Ciò si può anche giustificare: ne è testimone Georg Cantor, il cui lavoro sulla teoria degli insiemi e sui numeri trasfiniti, pur destinato a far epoca, fu messo in ridicolo dai principali matematici del suo tempo. Pare che tale esperienza avesse profondamente logorato i suoi nervi.

La maggior parte del lavoro di Gauss nel campo della teoria dei numeri riguardava il problema dei numeri com-



I numeri complessi della forma $a + bi$ (dove a e b sono interi e i è la radice quadrata di -1) si possono rappresentare come coppie di numeri (a, b), o punti di un piano, allo stesso modo in cui i numeri reali si rappresentano come punti di una retta. Le coppie di numeri complessi possono quindi essere trattate in modo geometrico. Per esempio, la rotazione di 90 gradi del segmento congiungente l'origine con il punto (a, b) equivale alla moltiplicazione per i del numero complesso originale. (Nella figura sono indicate tre di tali rotazioni.) Gauss fu il primo matematico a far notare che tale interpretazione geometrica poteva essere usata per ricavare in modo puramente algebrico le definizioni della somma e del prodotto.

plessi, definiti come numeri della forma $a+bi$, dove a e b sono interi e i è la radice quadrata di -1 (tale quindi che $i^2 = -1$). I numeri complessi furono introdotti dagli studiosi di algebra del Rinascimento, ma essi erano avvolti profondamente in un'aura di mistero. Perfino un uomo intelligente come Leibniz fu terribilmente confuso dall'argomento, giungendo a scrivere a proposito di questi misteriosi numeri che non sono né positivi né negativi: «Lo Spirito Divino ha trovato uno sbocco sublime in quella meraviglia dell'analisi, in quel portento del mondo ideale, in quello stato tra essere e non-essere, che noi chiamiamo radice immaginaria dell'unità negativa.»

Gauss fu più prosaico, preferendo usare una rappresentazione geometrica dei numeri complessi come punti di un piano. Pur essendo stata divulgata per la prima volta nel 1797, quando un geodeta norvegese, Caspar Wessel, pensò a una rappresentazione analitica della geometria del piano che era equivalente ai numeri complessi, il risultato di Gauss non divenne noto fino al 1897. Un contabile svizzero, Jean Robert Argand, sviluppò una descrizione simile nel 1806 e oggi la rappresentazione dei numeri complessi gli viene generalmente attribuita dai libri di testo. Il contributo di Gauss andava al di là di una definizone puramente geometrica dei numeri complessi. In una lettera scritta nel 1837 egli afferma di essersi accorto nel 1831 che si può evitare una interpretazione geometrica usando coppie della forma (a, b) al posto di $a+bi$, con definizioni puramente algebriche della somma e del prodotto. Usando coppie di numeri, egli mostrò che le operazioni aritmetiche con i numeri complessi sono definite dalle seguenti regole: $(a, b) + (c, d)$ è uguale ad $(a+c, b+d)$, mentre $(a, b)\,(c, d)$ è uguale ad $(ac-bd, ad+bc)$. È facile verificare che la coppia $(a, 0)$ si comporta proprio come il numero reale a e che $(0, 1)^2$ è uguale a $(-1, 0)$. Così la misteriosa radice quadrata di -1 viene a essere la coppia di numeri $(0, 1)$. La reale importanza di questo metodo, che è quello comunemente seguito oggi, non venne apprezzata fin quando William Rowan Hamilton lo pubblicò nel 1837. Gauss fu il primo matematico a fare libero ed esteso uso dei numeri complessi e ad accettarli pienamente come veri concetti matematici.

Nella sua tesi di laurea all'Università di Helmstedt, presentata nel 1799, Gauss diede la prima dimostrazione del «teorema fondamentale dell'algebra» (che oggi viene dimostrato in modo più naturale come teorema di topologia): ogni equazione algebrica ammette almeno una radice complessa. Anche in questo caso Gauss considerò il teorema talmente importante da dimostrarlo in quattro modi diversi durante la sua vita. La terza dimostrazione è particolarmente indicativa del suo stile impenetrabile e del suo talento originale. Egli costruisce a partire da un'equazione algebrica una complicata espressione sotto forma di un integrale doppio. Se l'equazione non ha radici, questo integrale deve assumere lo stesso valore sia integrandolo rispetto alla prima variabile e poi rispetto alla seconda che viceversa. Gauss dimostra che ciò non è vero e che i due procedimenti forniscono risultati differenti; perciò l'ipotesi che non vi siano radici è falsa e quindi esiste una radice.

Per capire da dove derivava la dimostrazione dobbiamo sapere che Gauss conosceva i teoremi fondamentali dell'analisi complessa, ma non li aveva pubblicati. La sua dimostrazione è una traduzione di un argomento della teoria delle funzioni di variabili complesse in termini più usuali di teoria delle funzioni di variabili reali. I risultati di questa traduzione sono rigorosi e logici, ma richiedono una certa complessità di formulazione che nasconde i principi fondamentali. Gauss aveva evidentemente capito che i suoi teoremi di analisi complessa non erano sufficientemente completi per poterli pubblicare e di conseguenza riformulò la sua dimostrazione.

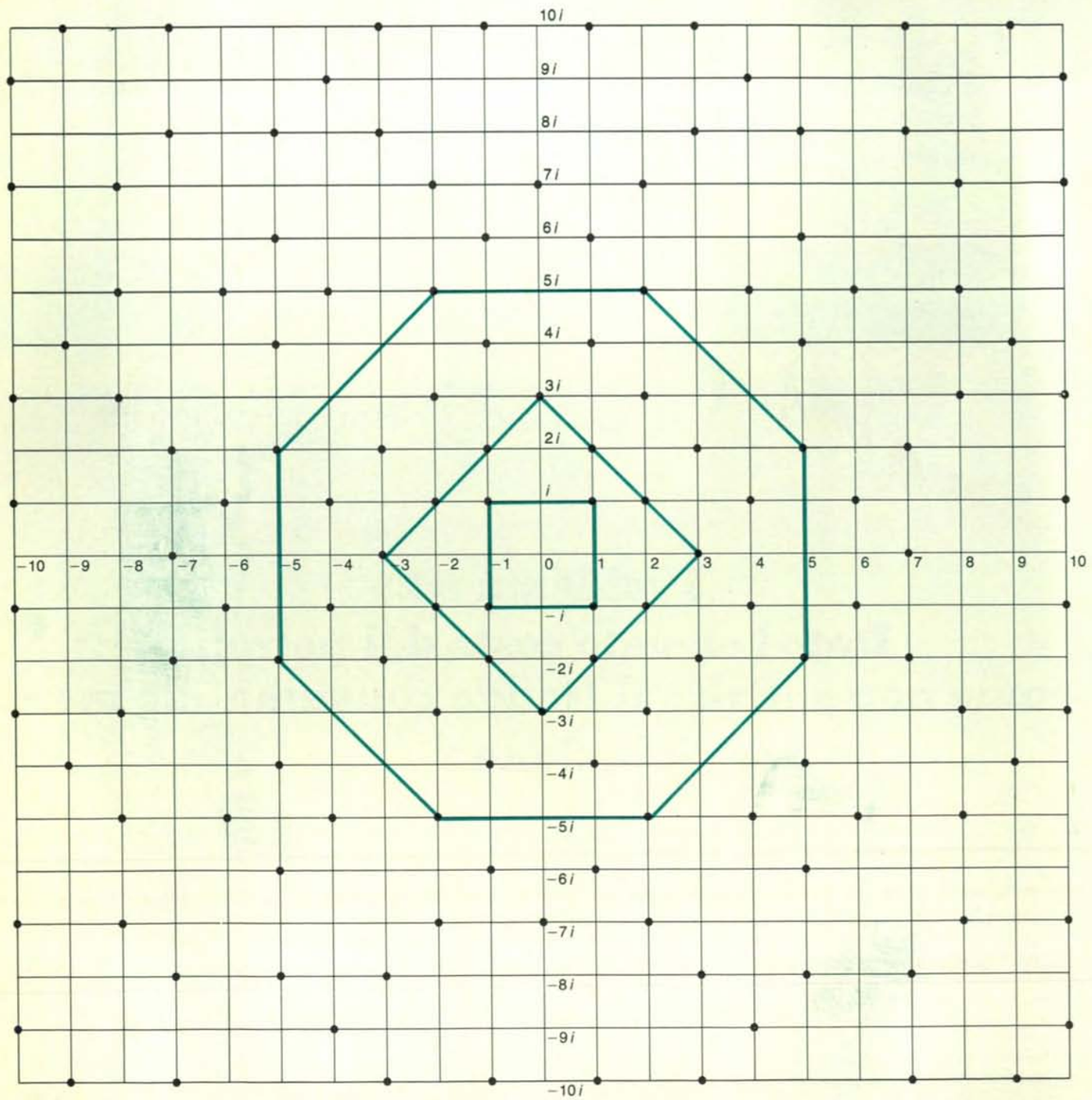
Gauss fece molto di più con i numeri complessi. Nel 1811 scoprì quello che oggi è chiamato teorema di Cauchy: l'integrale di una funzione analitica complessa lungo una linea chiusa che non comprende singolarità è nullo. Augustin Leom Cauchy lo considerò la base dell'analisi complessa, di cui è ancor oggi un punto fermo. Non lo chiamiamo teorema di Gauss perché Gauss non lo pubblicò mai. Sembra probabile che egli intendesse scrivere un lavoro definitivo sull'analisi dei numeri complessi, ma non trovò mai il tempo di elaborarlo in una forma che lo soddisfacesse.

Gauss sviluppò anche un metodo per fattorizzare numeri primi con numeri complessi ottenendo alcuni risultati sorprendenti. Per esempio, il numero primo 2 può essere fattorizzato nella forma $(1+i)\,(1-i)$. Allo stesso modo, 5 può essere fattorizzato come $(2+i)\,(2-i)$, 29 come $(5+2i)\,(5-2i)$ e così via. Certi numeri primi però (tra essi 7,11 e 19) non possono essere fattorizzati e rimangono primi. Gauss scoprì che soltanto i numeri primi della forma $4n+1$ possono essere fattorizzati con numeri complessi (il numero 2 è un caso particolare) e che ognuno di tali numeri primi può essere fattorizzato in un solo modo. Tali metodi vennero usati successivamente per risolvere problemi che in apparenza non

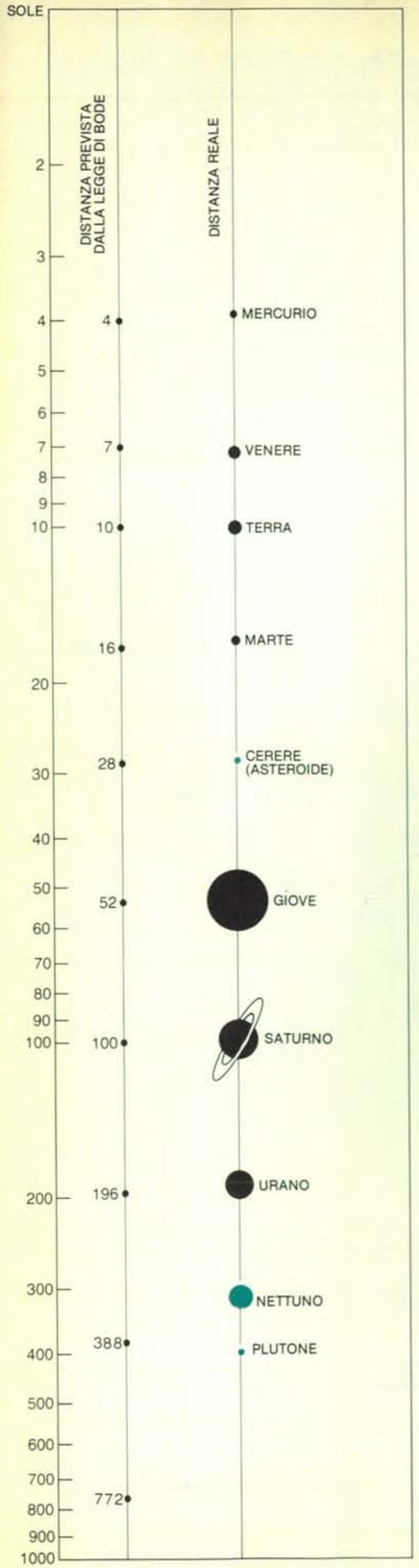
riguardavano affatto i numeri complessi.

In particolare Gauss fece uso dei numeri complessi della forma $a+bi$ (chiamati oggi interi di Gauss) per formulare e dimostrare una versione della legge di reciprocità quadratica per i residui biquadratici. Si dice che il numero k è un residuo biquadratico di un altro numero m se k è congruente modulo m alla quarta potenza di un intero. Così i residui biquadratici di 10 sono 0, 1, 5 e 6. La legge di reciprocità biquadratica afferma che per due numeri primi dispari p e q vi è un legame tra l'affermazione «p è un residuo biquadratico di q» e l'affermazione «q è un residuo biquadratico di p» purché siano soddisfatte un certo numero di condizioni sulla forma di p e di q. Questo teorema è analogo alla legge di reciprocità quadratica, ma risulta molto più complicato esprimerlo in termini matematici (ed è quindi molto difficile da impostare, non soltanto da dimostrare). Se si estende il teorema al caso in cui p e q sono interi di Gauss della forma $a+bi$, vi sono notevoli semplificazioni sia nell'enunciato del teorema che nella sua dimostrazione. Così il passaggio al caso dei numeri complessi fa sì che la trattazione del problema sia piú facile e anche più naturale che nel caso dei numeri reali.

La dimostrazione data da Gauss della reciprocità biquadratica mediante l'uso degli interi di Gauss ha costituito un procedimento standard per la soluzione di molti problemi della teoria dei numeri: in primo luogo si estende il teorema a un dominio di numeri complessi opportunamente scelti, chiamato campo numerico, dove è possibile una trattazione più naturale, quindi si risolve il problema in tale campo per poi ritornare al caso degli interi ordinari alla fine della dimostrazione. Questo efficace metodo ha spalancato le porte a quella che è attualmen-



I primi di Gauss, numeri complessi della forma $a + bi$ privi di fattori dello stesso tipo, sono distribuiti irregolarmente nel piano complesso. Gauss ne scoprì tre classi: (1) ± 1 $\pm i$, ai vertici di un quadrato, (2) $\pm p$ e $\pm pi$ (p è primo e uguale a $4n + 3$), formanti un rombo, e (3) $\pm a \pm bi$ e $\pm b \pm ai$, formanti un quadrato con gli angoli tronchi. I primi di Gauss formano sempre uno di questi schemi.



Le distanze dei pianeti dal Sole sono date approssimativamente dalla legge di Bode, scoperta per primo da Johann Titius nel 1766. (Ogni unità è uguale a circa 15 milioni di chilometri.) Il vuoto nella serie fece intensificare le ricerche del «pianeta mancante», che culminarono con la scoperta dell'asteroide Cerere nel 1801. Gauss si assunse il difficile compito di calcolare l'orbita di Cerere con i pochi dati disponibili. L'asteroide ricomparve dove egli aveva previsto. (I pianeti scoperti dopo il 1800 sono in colore. Si noti che la legge di Bode individua con scarsa approssimazione le posizioni dei pianeti più distanti di Urano).

te chiamata teoria dei numeri algebrica.

Nel 1801 gli interessi matematici di Gauss cambiarono bruscamente direzione allorché si dedicò all'astronomia. La sua passione per il calcolo ebbe sicuramente una parte importante. Nei suoi lavori più dotti vi sono lunghi calcoli e alcuni dei suoi più profondi teoremi di teoria dei numeri furono ricavati esaminando lunghi elenchi di cifre. Inoltre Gauss estese molti dei suoi calcoli fino a 21 decimali molto tempo prima dell'avvento di qualsiasi tipo di elaboratore.

L'interesse di Gauss per l'astronomia può farsi risalire a una scoperta di Johann Titius, che nel 1776 enunciò una legge empirica sulle distanze tra il Sole e i pianeti. Titius partì dalla successione 0, 3, 6, 12, 24, 48, e 96, nella quale ogni termine successivo al secondo è doppio di quello precedente, quindi aggiunse 4 a ogni termine per ottenere 4, 7, 10, 16, 28, 52 e 100. Tali numeri risultarono quasi esattamente proporzionali alle distanze dal Sole di Mercurio, Venere, Terra, Marte, Giove e Saturno; non vi era però alcun pianeta a distanza 28. Questa legge, che doveva diventare nota come legge di Bode (perché Johann Bode se la attribuì ingiustificatamente), rimase solo una curiosità fino al 1781, anno in cui William Herschel scoprì Urano a una distanza di circa 196 unità. Dato che il successivo termine della serie di Titius-Bode era $2(96)+4=196$, l'interesse era ora localizzato sul vuoto esistente alla distanza 28.

Nell'ultimo giorno dell'anno 1800 Giuseppe Piazzi scoprì quello che riteneva fosse il pianeta mancante. Si trattava di Cerere, che oggi sappiamo essere uno delle migliaia dei piccoli corpi della fascia degli asteroidi che si trova tra Marte e Giove. Una volta individuato Cerere, divenne importante calcolare l'orbita ellittica del nuovo corpo prima che gli osservatori ne perdessero le tracce. La difficoltà di osservazione di un oggetto così piccolo rese scarsi i dati disponibili ed estremamente difficile il calcolo dell'orbita: lo stesso Newton aveva fatto notare quanto fosse difficile calcolare le orbite con pochi dati.

Dopo solo tre osservazioni Gauss mise a punto una tecnica per calcolare le componenti orbitali con una precisione tale che tra la fine del 1801 e l'inizio del 1802 parecchi astronomi furono in grado di ritrovare Cerere senza difficoltà. Tra l'altro egli mostrò che le variazioni relative alle informazioni ricavate sperimentalmente potevano essere rappresentate da una curva a campana (meglio conosciuta oggi come distribuzione gaussiana). Egli mise a punto anche il metodo dei minimi quadrati, mediante il quale il valore più attendibile viene ricavato con un particolare calcolo che minimizza le somme di differenze al quadrato. I suoi metodi, descritti in un suo articolo del 1809 dal titolo *Teoria del moto dei corpi celesti che ruotano attorno al Sole su sezioni coniche* sono ancora impiegati oggi. Sono necessarie soltanto poche modifiche per renderli adatti ai moderni elaboratori elettronici.

Gauss ebbe lo stesso successo nel determinare l'orbita dell'asteroide Pallade, per il quale perfezionò i suoi calcoli per tener conto delle perturbazioni prodotte dagli altri pianeti del sistema solare. Nel 1807 divenne professore di astronomia e direttore del nuovo osservatorio dell'Università di Göttingen, dove rimase per il resto della sua vita. La sua prima moglie morì nel 1809, subito dopo la nascita del loro terzo figlio. Dal secondo matrimonio nacquero due figli e una figlia.

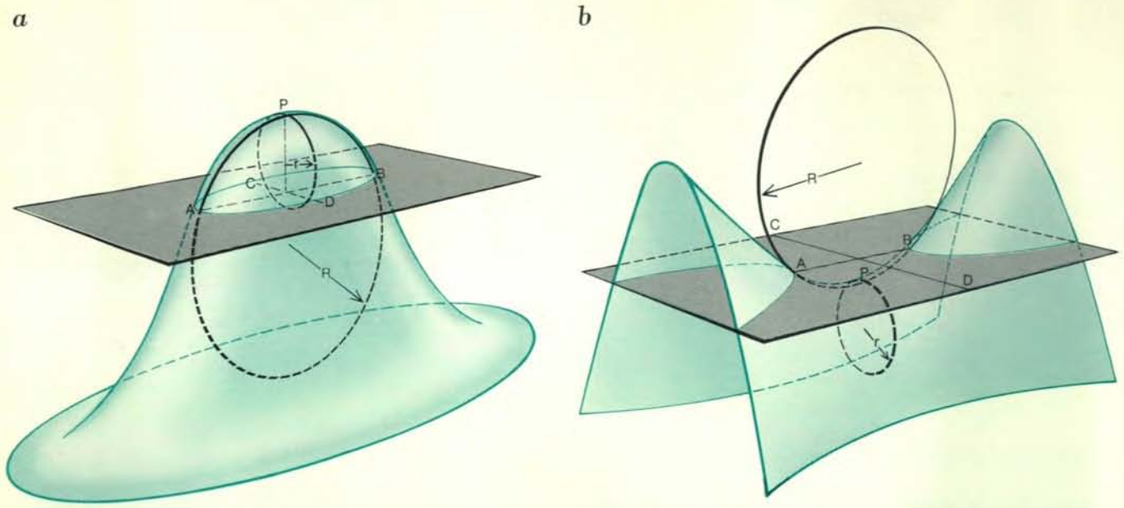
Verso il 1820 Gauss rivolse la sua attenzione alla geodesia: la determinazione matematica della forma e delle dimensioni della Terra. A tale scienza dedicò la maggior parte dei successivi otto anni in studi teorici e di ricerche empiriche. Nel 1821 divenne consigliere scientifico dei governi di Hannover e di Danimarca, che gli affidarono l'incarico di redigere una mappa geodetica di Hannover mediante la tecnica della triangolazione. A tale scopo mise a punto l'eliotropo, uno strumento che rifletteva i raggi del Sole in una direzione ben determinata, rendendo in tal modo possibile l'allineamento di strumenti di rilevamento su lunghe distanze.

Gli sforzi di Gauss per determinare la forma della Terra con effettive misurazioni geodetiche lo riportarono alla pura teoria. Lavorando con i dati delle sue esplorazioni, egli sviluppò una teoria delle superfici curve in cui le caratteristiche di una superficie possono essere determinate semplicemente misurando la lunghezza delle curve giacenti su di essa. Questa teoria della superficie intrinseca ispirò a uno dei suoi allievi, Bernhard Riemann, lo sviluppo di una geometria generale intrinseca degli spazi a tre o più dimensioni. Circa 60 anni dopo le idee di Riemann costituirono le basi matematiche della teoria generale della relatività di Albert Einstein.

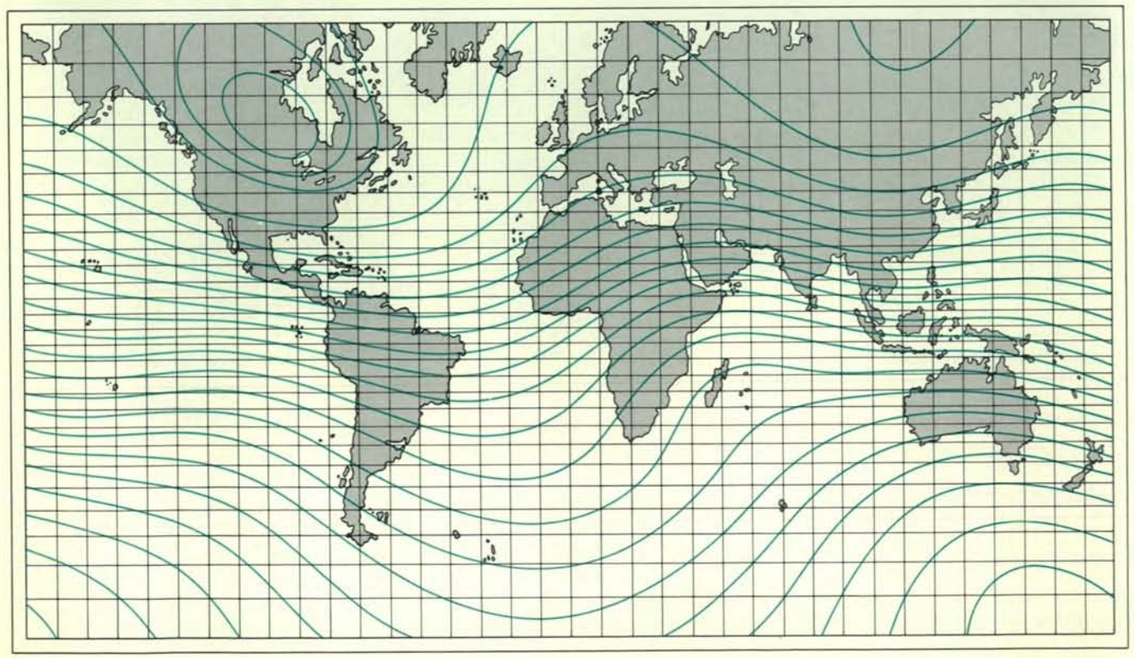
A partire dal 1831, quando arrivò a Göttingen il fisico Wilhelm Weber, Gauss lavorò in stretta collaborazione con lui a ricerche sperimentali e teoriche sul magnetismo. Insieme inventarono un magnetometro e come risultato del loro interesse per il magnetismo terrestre organizzarono una rete di osservatori attraverso l'Europa per misurare le variazioni del campo magnetico terrestre. Gauss poté dimostrare con un'analisi teorica che

il campo nasceva all'interno della parte solida della Terra, risultato di considerevole importanza in quanto limitava le possibili origini del campo e focalizzava l'attenzione sui meccanismi geofisici della sua produzione. Il suo contributo è oggi riconosciuto con il gauss, l'unità di densità di flusso magnetico nel sistema di unità di misura di Gauss.

Gauss e Weber furono anche tra i primi a sottolineare la possibilità di inviare messaggi con l'elettricità. La storia del telegrafo è lunga, ma fino ai primi anni dell'ottocento si faceva uso soltanto di metodi non elettrici. Quando poi nel 1827 venne trasmesso un impulso elettrico per una distanza di circa 300 metri, nacque immediatamente l'idea di poter



La teoria della superficie intrinseca sviluppata da Gauss permette di calcolare la curvatura di una superficie misurando la lunghezza delle curve giacenti su tale superficie. La curvatura di una superficie convessa (*a*) si trova così: si tracci in un primo luogo un piano, parallelo al piano tangente nel punto *P*, che interseca la superficie nei pressi di *P* lungo un'ellisse (*colore scuro*). Si traccino gli assi maggiore e minore dell'ellisse, indicati con *AB* e *CD* sulla superficie per ricavare le curve *APB* e *CPD*. Si trovino successivamente due cerchi che si adattino il più possibile alle curve proiettate nel punto *P*. Se i raggi dei due cerchi sono *R* e *r*, quando il piano si avvicina a *P* e l'ellisse diventa molto piccola, la curvatura si avvicina a $1/Rr$. In (*b*) la superficie è a forma di sella e il piano la interseca lungo due iperboli. Per convenzione un raggio (e quindi la curvatura) è negativo. Nello spazio piano *R* ed *r* hanno lunghezza infinita e la curvatura è nulla.



Una mappa del campo magnetico terrestre ricavata da un disegno apparso nel 1840 in *Atlas des Erdmagnetismus*, scritto in collaborazione da Gauss e Weber. I due scienziati ottennero le misure sperimentali per la mappa organizzando una rete di osservatori in tutto il mondo.

utilizzare l'elettricità in telegrafia. Furono progettati vari telegrafi elettrici, ma nessuno di essi fu realizzato fino al 1832, quando furono collegate le residenze estiva e invernale dello Zar a Pietroburgo. Un anno dopo Gauss e Weber misero a punto un telegrafo che correva sopra i tetti di Göttingen per una distanza di 2,3 chilometri. I segnali venivano trasmessi come una sequenza di cinque deflessioni di un ago verso destra o verso sinistra (con un totale di 32 possibilità) e il dispositivo funzionava così bene che i due uomini lo usavano regolarmente per comunicare tra loro. Il telegrafo di Gauss-Weber fu probabilmente il primo a funzionare con una certa utilità e precedette di sette anni il famoso brevetto di Samuel F.B. Morse.

La grande fama di Gauss fu ulteriormente arricchita dalla scoperta di risultati non pubblicati che anticipavano molti dei principali progressi del secolo XIX. Oltre al teorema di Cauchy egli aveva scoperto la doppia periodicità delle funzioni ellittiche, che per merito di Jacobi e Abel diventarono il cardine della teoria delle funzioni del secolo XIX. Le funzioni ellittiche sono un tipo particolare di funzioni $f(z)$ di un numero complesso z. Doppia periodicità vuol dire che esistono due differenti costanti complesse, per esempio a e b, tali che, per qualsiasi z, $f(z)$ è uguale a $f(z+a)$, che a sua volta è uguale a $f(z+b)$. C'è una certa analogia con la periodicità semplice delle funzioni trigonometriche, per le quali $\sin(z)$ gode della proprietà che $\sin(z + 2\pi) = \sin(z)$. La scoperta di Gauss ha notevoli conseguenze per quanto concerne le relazioni esistenti tra la teoria dei numeri e la teoria delle funzioni complesse.

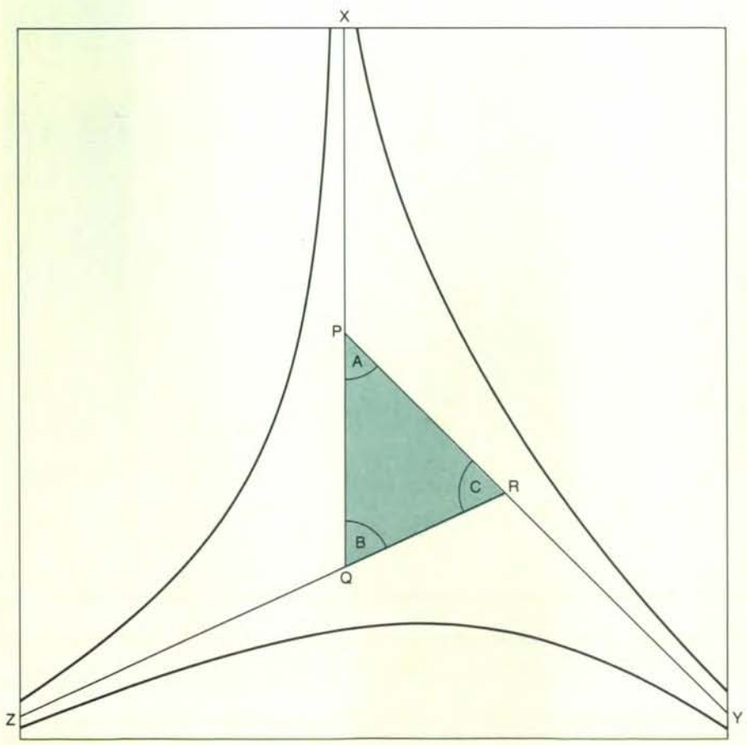
Gauss fu anche tra i primi a mettere in dubbio che la geometria euclidea fosse innata nella natura e nel pensiero umano. La geometria di Euclide era basata su certi assiomi, o proposizioni fondamentali, che erano considerati come verità autoevidenti. L'intero sistema era costruito per via puramente deduttiva a partire da tali basi non dimostrate. L'assioma delle parallele afferma che dati una linea retta e un punto fuori di essa, si può tracciare per questo punto soltanto una retta parallela alla retta data. L'assioma delle parallele fu intensamente studiato perché, diversamente dagli altri assiomi, non era a prima vista convincente e richiedeva il concetto di infinito. Fin dalla antichità i matematici avevano cercato invano di sostituirlo con assiomi ritenuti più evidenti.

Nel secolo XVIII vi fu un rinnovato interesse per questo problema insoluto e molti matematici, professionisti e dilettanti tentarono di dimostrare che l'assioma delle parallele era una conseguenza logica degli altri assiomi di Euclide. Tutte queste dimostrazioni risultarono errate. Gauss venne a conoscenza della controversia mentre studiava a Göttingen e nel 1804 scrisse una lettera al matematico ungherese Farkas Bolyai in cui mostrava l'errore contenuto nella dimostrazione di Bolyai dell'assioma delle parallele, cioè la sostituzione di un argomento infinito con uno finito. Gauss aggiungeva nella stessa lettera che anche lui si era trovato nei guai per la stessa difficoltà. Nel 1815, però, recensendo alcuni libri, alludeva alla possibilità di una geometria per la quale l'assioma delle parallele non fosse valido, essendo tale geometria autoconsistente ed esente da contraddizioni. Conoscendo l'abituale cautela di Gauss nel rivelare le proprie idee, questa affermazione ci fa pensare seriamente che egli avesse qualche prova per sostenerla. Forse, visto che le sue idee si scontravano con quelle dei suoi contemporanei decise di non renderle pubbliche, avendo intuito, probabilmente a ragione, che non sarebbero state capite.

Nel 1820 il figlio di Bolyai, János, contagiato dalla fanatica preoccupazione del padre di dimostrare l'assioma del parallelismo, concluse che era impossibile dimostrarlo e cominciò a sviluppare una nuova geometria che non si basava sull'assioma di Euclide. Tre anni dopo portò a termine un articolo nel quale proponeva un sistema coerente di geometria non euclidea che venne pubblicato come appendice al libro del padre *Saggio sugli elementi di matematica per i giovani studenti*. Gauss lesse tale articolo nel 1832 e scrisse al vecchio Bolyai che non lo poteva approvare perché il farlo equivaleva ad approvare il lavoro che aveva fatto lui stesso 30 anni prima. Il giovane Bolyai fu profondamente amareggiato della critica di Gauss e morì praticamente senza alcun riconoscimento della sua soluzione a un problema così importante e annoso (risolto indipendentemente, all'incirca nello stesso periodo e quasi nello stesso modo, da Nikolai Ivanovich Lobachevsky). L'atteggiamento di Gauss fu del tutto scorretto dato che egli non si era mai sentito abbastanza sicuro del proprio lavoro per renderlo pubblico. Forse era un po' geloso del successo di Bolyai.

Per diverse ragioni Gauss era a un bivio. Egli può essere considerato sia come il primo dei matematici moderni sia come l'ultimo dei grandi matematici classici. Infatti i suoi metodi erano moderni nello spirito, ma la scelta degli argomenti era classica.

Una caratteristica dell'opera di Gauss, specialmente nella matematica pura, è il suo modo di ragionare sul particolare



La geometria iperbolica è un sistema non euclideo studiato da Gauss in cui è possibile trovare l'area di un triangolo noti che siano i suoi angoli, il che non vale nella geometria euclidea. Il grafico qui riportato fa parte della dimostrazione data da Gauss che l'area del triangolo è proporzionale alla differenza tra 180 gradi e la somma dei suoi angoli *A*, *B* e *C*. Dato che il grafico è tracciato su un piano euclideo le rette appaiono curvate, ma nello spazio iperbolico sono «rette». Inoltre le linee concorrenti a gruppi di tre agli estremi del grafico sono parallele, in modo tale che la linea apparentemente curva *XY* è parallela sia a *XQ* sia a *PY*, il che non accade nella geometria euclidea. Il triangolo *XYZ*, pur avendo i vertici all'infinito, ha area finita.

Anioni dei metalli alcalini

I metalli alcalini danno luogo a cationi, cioè a ioni con carica positiva dovuta alla perdita di un elettrone. Si è ora scoperto che acquistando un elettrone questi metalli possono dare luogo anche ad anioni molto reattivi

di James L. Dye

I metalli alcalini sono un gruppo di elementi le cui proprietà chimiche più notevoli e familiari derivano dalla facilità con cui perdono un elettrone. Il metallo alcalino sodio, per esempio, dona facilmente un elettrone al cloro, formando il cloruro di sodio. Il sodio metallico cede con tale facilità uno dei suoi elettroni che riesce addirittura a scindere la molecola dell'acqua, spostando un atomo di idrogeno e formando idrossido di sodio, reazione che può essere decisamente violenta. A causa della sua reattività, il sodio metallico non esiste in natura. Questo elemento si trova solo sotto forma di ione positivo, o catione, denotato da Na^+. Il catione si forma quando l'atomo dell'elemento cede uno dei suoi elettroni a qualche altra specie chimica.

Queste proprietà dei metalli alcalini sono note fin dagli albori della chimica e quindi è stata una vera sorpresa la scoperta che essi possono anche accettare un elettrone, comportandosi in maniera esattamente opposta a quella usuale. La aggiunta di un elettrone, per esempio, a un atomo neutro di sodio porta alla formazione di uno ione carico negativamente, o anione, denotato da Na^-. Da qualche tempo è noto che questo anione è una specie stabile presente nei vapori di sodio, ma, quel che è più importante, è stato recentemente scoperto in soluzione e anche come componente di un sale cristallino. È ormai chiaro che è possibile preparare anche gli anioni degli altri metalli alcalini; sono già stati osservati infatti quelli del potassio (K^-), del rubidio (Rb^-) e del cesio (Cs^-). Potrebbe anche essere possibile preparare sali contenenti il più semplice anione immaginabile: un elettrone immobilizzato.

Il punto chiave nella preparazione degli anioni dei metalli alcalini è l'intrappolamento dei cationi in una molecola organica dalla struttura a gabbia. In condizioni ordinarie qualsiasi anione metallico eventualmente presente in soluzione reagirebbe immediatamente con il corrispondente catione per formare due atomi neutri del metallo. Quando i cationi sono sequestrati dalla molecola che li intrappola, il complesso risultante è talmente stabile che la reazione suddetta non può avvenire. In ogni caso gli ioni negativi e i cristalli che li contengono sono estremamente reattivi. Non possono venire esposti all'aria o all'umidità e rimangono stabili per lunghi periodi di tempo solo a bassa temperatura.

Il gruppo dei metalli alcalini comprende sei elementi: litio, sodio, potassio, rubidio, cesio e francio. L'ultimo è un elemento radioattivo, molto raro. Nell'ordinaria disposizione della tavola periodica degli elementi, i metalli alcalini vengono collocati nella prima colonna, sotto l'idrogeno. Questo modo di raggrupparli riflette le somiglianze nel comportamento chimico di questa famiglia di metalli e quindi nella loro struttura elettronica: tutti, infatti, possiedono un solo elettrone nello strato elettronico più esterno, detto di valenza.

In qualsiasi atomo neutro il numero degli elettroni deve essere tale da bilanciare la carica elettrica dei protoni presenti nel nucleo. Gli elettroni tendono spontaneamente a disporsi secondo una configurazione di energia minima, ma nel far ciò devono rispettare i limiti imposti dalle regole della meccanica quantistica, che confinano gli elettroni nei cosiddetti orbitali. Dal punto di vista della fisica classica si può dire che un orbitale descrive la traiettoria seguita dall'elettrone intorno al nucleo. Un'interpretazione più corretta, dal punto di vista della meccanica quantistica, vede l'orbitale come la definizione della probabilità di trovare l'elettrone in un dato punto dello spazio. Si può quindi dire che l'orbitale rappresenta la distribuzione nello spazio della carica elettronica.

L'ordine di riempimento degli orbitali disponibili è determinato da una regola di esclusione che proibisce a due elettroni di occupare il medesimo stato energetico. Al primo orbitale da riempire viene dato il nome 1*s*. Gli elettroni dell'orbitale 1*s* sono distribuiti in una regione sferica con centro in corrispondenza del nucleo. Tale distribuzione è assolutamente simmetrica e non definisce alcuna direzione privilegiata. L'orbitale 1*s* può essere occupato da uno o due elettroni Nel secondo caso i due elettroni devono avere valori opposti dello spin o momento angolare intrinseco. In meccanica quantistica, infatti, l'elettrone ruota attorno al suo asse e tale rotazione può essere orientata nello spazio in due soli modi.

Il secondo orbitale da riempire viene denominato 2*s*. Ha forma simile all'1*s*, ma i due elettroni che lo riempiono sono dotati di maggiore energia e si dispongono quindi entro una sfera più grande.

A questo punto del procedimento di riempimento degli orbitali di un atomo si incontrano i primi orbitali non sferici. Sono detti orbitali *p* e ve ne sono tre, tutti designati con la sigla 2*p*. Ognuno dei tre orbitali 2*p* racchiude una regione dello spazio formata da due lobi, separati dal nucleo, e sono disposti nello spazio in modo che gli assi dei lobi risultino perpendicolari fra di loro. I sei elettroni che riempiono gli orbitali 2*p* (a coppie con spin opposti) in aggiunta al momento angolare intrinseco possiedono pure un momento angolare orbitale che dipende dal moto di rivoluzione. In altre parole, oltre a ruotare su sé stessi orbitano intorno al nucleo.

Per riempire gli orbitali 1*s*, 2*s* e 2*p* sono necessari 10 elettroni. Proseguendo nell'aggiunta, i nuovi elettroni riempiono dapprima l'orbitale 3*s* e successivamente tre orbitali 3*p*, sempre di forma simile agli orbitali *s* e *p* di minore energia, ma più grandi. Gli atomi più pesanti possiedono ulteriori tipi di orbitali, detti *d*, *f* e *g*, con geometrie più complicate e valori più elevati del momento angolare orbitale.

È di importanza fondamentale il fatto che un atomo privo di posti liberi in tutti i suoi orbitali *s* e *p* è caratterizzato da una particolare stabilità. Gli elementi con atomi siffatti costituiscono la famiglia dei gas nobili, la cui inerzia chimica testimonia la loro stabilità. Il primo di questi è l'elio, con appena due elettroni che riempiono completamente l'orbitale 1*s*. Il neo, con 10 elettroni, ha gli orbitali 1*s*, 2*s* e 2*p* completati; l'argo ha tutti gli orbitali riempiti fino a 3*s* e 3*p*. Tutti questi elementi cedono o accettano elettroni con difficoltà estrema.

I metalli alcalini possono essere considerati come dei gas nobili con un elettrone in più, cui corrisponde ovviamente un protone in più nel nucleo. Il litio possiede l'orbitale 1*s* completo del gas elio e in più un elettrone che riempie a metà l'orbitale successivo 2*s*. Il sodio, come il neo, ha gli orbitali 1*s*, 2*s* e 2*p* completi e in più un elettrone 3*s*. I rimanenti metalli alcalini hanno struttura elettronica simile: in tutti, infatti, si ha nella posizione più esterna un singolo elettrone in un orbitale *s*.

Si può capire ora la forte tendenza dei metalli alcalini a perdere un elettrone per formare un catione. Con la perdita dell'elettrone di valenza tutti i metalli alcalini assumono la struttura elettronica eccezionalmente stabile di un gas nobile. Anche se possiede una carica positiva non bilanciata, il catione risultante, in presenza di un accettore di elettroni, è più stabile dell'atomo neutro.

L'aggiunta di un elettrone a un metallo alcalino porta a una struttura elettronica molto meno favorita di quella tipo gas nobile del catione. Ciononostante l'elettrone in più riempie un orbitale *s* e un orbitale completo è più stabile di uno pieno a metà. In situazioni controllate con cura questo leggero aumento di stabilità è sufficiente per consentire l'esistenza di anioni dei metalli alcalini.

La chimica delle specie atomiche ionizzate si studia più facilmente considerando le soluzioni. Quindi per studiare i metalli alcalini in soluzione occorre un solvente capace di scioglierli. Il filone di ricerca che ha portato recentemente all'osservazione degli anioni dei metalli alcalini ha avuto inizio in Germania nel 1864 con la scoperta da parte di W. Weyl che i metalli alcalini si sciolgono in ammoniaca liquida anidra (NH_3). Le soluzioni risultanti hanno un bel colore blu molto intenso.

Il processo di soluzione di una sostanza ionica è la conseguenza dell'interazione elettrica fra le molecole del solvente e gli ioni del solido. In generale il solvente deve essere un liquido polare, deve cioè essere costituito da molecole che, sebbene complessivamente neutre, hanno una distribuzione di carica asimmetrica. La ammoniaca è un esempio di molecola polare. Nello stato liquido la molecola ha una configurazione per cui i tre atomi di idrogeno sono raggruppati l'uno vicino all'altro tutti da una parte dell'atomo di azoto. Questo ha un eccesso di carica negativa, bilanciato dalla carica positiva degli atomi di idrogeno. Ne segue che la molecola è un dipolo elettrico, con cariche uguali e contrarie separate da una piccola distanza.

Quando si scioglie un sale come il cloruro di sodio, la molecola si spezza in

Gli anioni del sodio sono un costituente essenziale del liquido blu e del film dorato formatisi entro l'apparecchio a vuoto in vetro. Per la preparazione si è disposto dapprima uno strato di sodio metallico dentro il pallone a sinistra. Parte del sodio è stato poi disciolto in etilammina. Nella soluzione gli ioni positivi del sodio (cationi) sono intrappolati entro le molecole di una sostanza organica, detta criptante, che ha una struttura a gabbia. Gli ioni negativi (anioni) sono liberi in soluzione e le impartiscono il colore blu intenso visibile sul fondo a destra. Il film, che si forma sulle pareti del recipiente quando evapora una parte del solvente, è costituito da anioni e cationi intrappolati. La manopola di plastica in alto appartiene a una valvola per vuoto.

I cristalli di questo sale sono composti da anioni sodio e da un complesso di cationi sodio e criptante. Come tutti i sali, questo si forma quando una specie chimica dona un elettrone a un'altra specie, formando una coppia di ioni. Il fatto inconsueto in questo caso è che il donatore e l'accettore di elettroni sono due forme dello stesso elemento. I cristalli sono stati precipitati da una soluzione per raffreddamento. La loro brillantezza metallica è ingannevole: il sale non ha comportamento metallico ma è un semiconduttore. Il sale non può essere esposto all'aria e anche sotto vuoto si decompone lentamente se non viene mantenuto a bassa temperatura.

ioni positivi e negativi, nel nostro caso Na^+ e Cl^-. Quello che tiene in soluzione tali ioni è l'interazione con il solvente. La carica positiva dello ione Na^+, per esempio, lo fa attrarre dalla regione più negativa del solvente, con la conseguenza che lo ione viene presto circondato da uno strato di molecole di solvente orientate. L'anione Cl^- attrae esso pure uno strato di molecole di solvente, che sono però orientate in senso opposto. Sono questi strati di molecole di solvente orientate che stabilizzano la soluzione e impediscono agli ioni di aggregarsi formando un sale.

La maggior parte dei metalli sono molto meno solubili dei sali semplicemente perché non si forma facilmente un anione. I metalli alcalini si sciolgono in ammoniaca a causa della facilità con cui cedono un elettrone. La soluzione viene così a contenere cationi metallici indistinguibili da quelli risultanti dalla dissoluzione di un sale, ma gli ioni negativi sono fuori del comune: si tratta semplicemente di elettroni liberati nella soluzione. Essi vengono detti elettroni solvatati e si comportano in maniera molto simile a qualsiasi altra specie chimica negativa, attraendo e orientando le molecole polari del solvente, ma, dato che sono migliaia di volte più leggeri di qualunque altro ione, sono notevolmente più mobili.

Come è possibile identificare le specie chimiche presenti in soluzione? Come può essere dimostrato, per esempio, che le soluzioni dei metalli alcalini in ammoniaca contengono elettroni solvatati e non qualche altro ione negativo, magari proprio anioni metallici? Questi problemi possono venire affrontati con svariate tecniche, ma ci occuperemo in questa sede solamente di due. Con queste metodiche gli ioni in soluzione vengono caratterizzati in base alle loro proprietà magnetiche e in base alla lunghezza d'onda della radiazione da essi assorbita.

Il magnetismo degli elettroni deriva, in ultima analisi, dal loro spin. La rotazione di un singolo elettrone origina un piccolo campo magnetico praticamente allo stesso modo in cui viene originato da un flusso di elettroni circolanti in una spira metallica. Questi campi possono essere osservati solo in pochi atomi. Occorre ricordare che in un orbitale completo vi sono due elettroni accoppiati con spin opposti, quindi i due campi magnetici da essi prodotti si annullano a vicenda. In un atomo privo di orbitali incompleti tutti i campi si annullano e il magnetismo complessivo è zero. Una sostanza fatta di atomi di questo tipo viene detta diamagnetica: non ha un campo magnetico proprio e quando è immersa in un campo esterno gli si oppone debolmente, provocando un allargamento delle linee di forza.

Gli atomi con orbitali incompleti e quindi con spin elettronici non accoppiati possono esibire un campo magnetico non nullo. Se questo campo è molto forte, come nel ferro e in altri metalli a esso molto simili, il materiale viene detto ferromagnetico. In tali materiali gli spin degli elettroni disaccoppiati si allineano spontaneamente, almeno in domini microscopici, provocando un magnetismo permanente. Nella maggior parte delle sostanze magnetiche, ivi compresi i metalli alcalini, il campo associato a ogni atomo è molto più piccolo, troppo piccolo perché possa aversi un allineamento spontaneo degli spin. Un allineamento è comunque possibile, con la conseguenza di una leggera magnetizzazione della sostanza, in seguito a immersione in un forte campo magnetico esterno. I materiali che possiedono queste proprietà vengono detti paramagnetici.

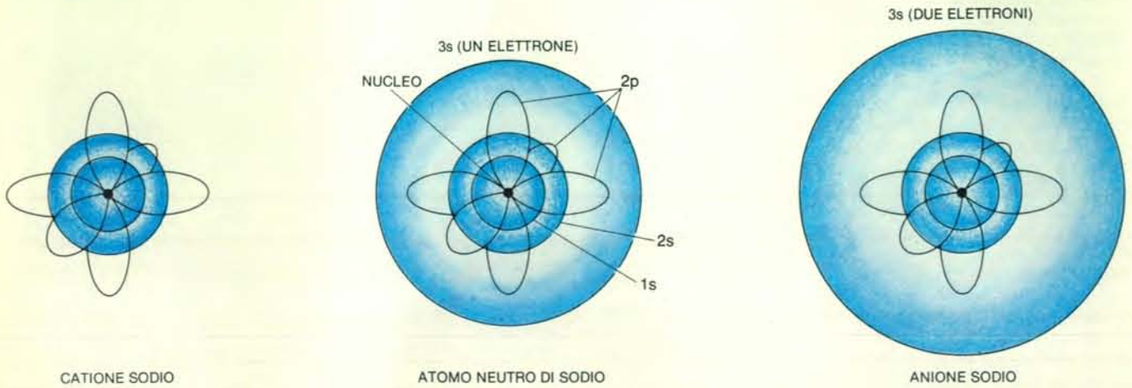
Gli elettroni solvatati in una soluzione di metallo alcalino in ammoniaca dovrebbero comportarsi come elettroni spaiati in un atomo e quindi rendere la soluzione paramagnetica. Inoltre l'entità del paramagnetismo dovrebbe essere proporzionale alla concentrazione degli elettroni solvatati. Soluzioni del genere sono risultate in effetti paramagnetiche e non vi è dubbio che contengano elettroni solvatati. Però l'intensità del fenomeno osservato non è proporzionale alla concentrazione, con la conseguenza che devono essere presenti anche altre specie ioniche. La natura delle deviazioni dalla proporzionalità indicano che alcuni elettroni si accoppiano per formare una specie diamagnetica.

La natura dello ione diamagnetico nelle soluzioni ammoniacali dei metalli alcalini è stata oggetto di controversia per decenni, e ancora oggi il problema non è del tutto appianato. Un candidato ovvio è l'anione metallico, dato che un metallo alcalino che acquista un elettrone ha tutti gli orbitali completi e deve quindi essere diamagnetico. La prima ipotesi di questo tipo venne proposta nel 1953 dal ricercatore tedesco Werner Bingel, ma la

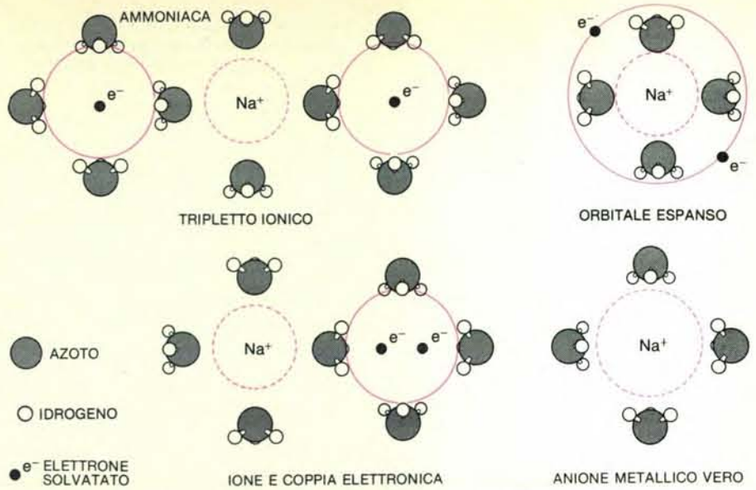
H IDROGENO																	He ELIO
Li LITIO	Be											B	C	N	O	F	Ne NEO
Na SODIO	Mg											Al	Si	P	S	Cl	Ar ARGO
K POTASSIO	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr KRIPTO
Rb RUBIDIO	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe XENO
Cs CESIO	Ba	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn RADO
Fr FRANCIO	Ra	Ac															

METALLI ALCALINI — GAS NOBILI

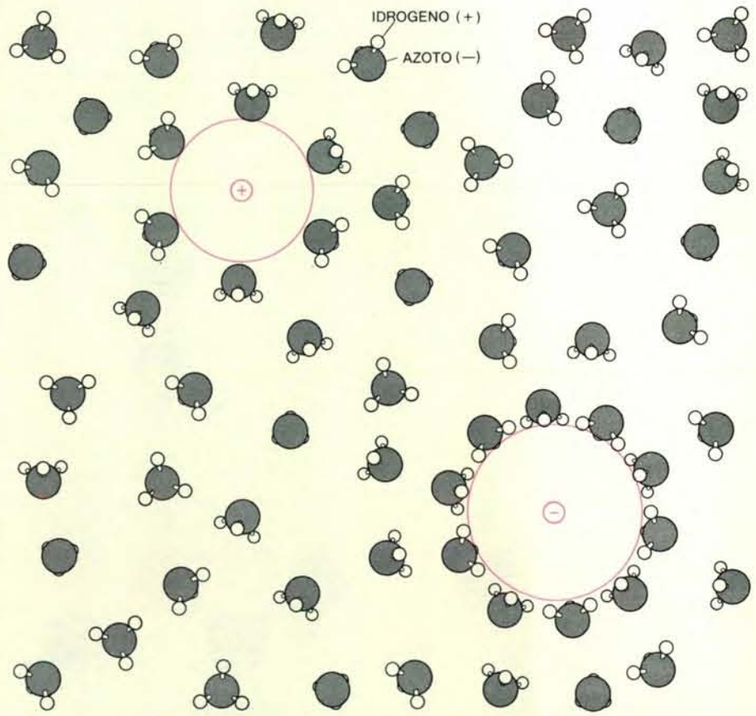
I metalli alcalini sono gli elementi elencati sotto l'idrogeno nella prima colonna del sistema periodico. Essi vengono raggruppati a causa della somiglianza della loro struttura elettronica e quindi delle loro proprietà chimiche. Dal punto di vista della configurazione elettronica, i vari metalli alcalini possono venire considerati come gas nobili cui sia stato aggiunto un elettrone. I gas nobili, elencati all'estrema destra, sono gli elementi più stabili e i metalli alcalini cedono con molta facilità un elettrone per formare un catione con la struttura elettronica di un gas nobile. L'elettrone in più viene eliminato con tale facilità che in certi casi può essere donato a un altro atomo di metallo alcalino.



L'acquisto o la donazione di un elettrone modifica profondamente le proprietà chimiche dei metalli alcalini. L'atomo di sodio neutro (*al centro*) possiede due elettroni in ognuno dei cinque orbitali interni e uno solo in quello più esterno o di valenza. Gli orbitali definiscono la distribuzione spaziale della carica elettronica e non possono contenere più di due elettroni ciascuno. L'orbitale di minore energia, detto 1*s*, ha una forma sferica. Anche il secondo, detto 2*s*, è sferico, ma più grande. Seguono tre orbitali 2*p*, costituiti da due lobi. Infine l'orbitale più esterno 3*s* è ancora sferico. Nel catione (*a sinistra*) l'elettrone 3*s* è stato ceduto e lo ione assume la configurazione elettronica del neo, con orbitali 1*s*, 2*s* e 2*p* completi. Nell'anione sodio (*a destra*) l'elettrone in eccedenza è andato a occupare il posto libero nell'orbitale 3*s*.



Oltre ai veri anioni dei metalli alcalini, nelle soluzioni di questi ultimi esistono altri possibili ioni negativi. Sono qui raffigurati schematicamente alcuni di questi ioni ipotizzabili nelle soluzioni di sodio in ammoniaca. Un catione sodio (Na^+) può attirare nelle sue vicinanze due elettroni solvatati indipendenti, formando una tripletta ionica. Oppure il catione può attirare una coppia elettronica, formata da due elettroni circondati da una gabbia di molecole di solvente. Infine due elettroni possono associarsi a un catione in un orbitale espanso che, oltre a contenere il catione, comprende alcune molecole di solvente. Tutte queste possibili strutture sono caratterizzate dal fatto che fra il catione e gli elettroni più esterni sono comprese delle molecole di solvente, mentre nel vero anione del metallo alcalino i due elettroni occupano un comune orbitale 3*s*.



Sono le interazioni elettrostatiche fra le molecole del solvente e gli ioni in soluzione che impediscono a questi ultimi di riaggregarsi e di precipitare. Solo i liquidi con molecole polari sono dei solventi efficaci per i metalli alcalini. L'ammoniaca ne è un esempio: l'atomo di azoto nell'ammoniaca liquida tende ad acquistare una carica negativa che viene bilanciata dalla carica positiva degli atomi di idrogeno. Nei liquidi polari gli ioni attirano uno strato di molecole di solvente che li isola da quelli con carica opposta. L'ammoniaca scioglie tutti i metalli alcalini.

specie che questi aveva in mente non era proprio quella di un vero anione metallico. Egli immaginava che un comune catione potesse attrarre nei suoi pressi una coppia di elettroni legati fra di loro. La possibilità che le soluzioni ammoniacali potessero contenere dei veri anioni metallici venne discussa la prima volta nel 1965 da Thomas R. Tuttle, Jr., Sidney Golden e Charles Guttman della Brandeis University.

Anioni metallici possono benissimo esistere nelle soluzioni ammoniacali dei metalli alcalini, ma è onesto dire che non vi è una prova evidente della loro presenza. Inoltre lo spettro di assorbimento delle soluzioni indica, anche se un po' indirettamente, che l'anione non può essere la sola specie diamagnetica presente.

Il colore blu intenso delle soluzioni ammoniacali dei metalli alcalini è dovuto a un forte assorbimento della luce all'estremo rosso dello spettro visibile e nella regione infrarossa. Questo assorbimento della radiazione di grande lunghezza di onda è caratteristico nelle sostanze contenenti elettroni non strettamente legati a singoli atomi. I metalli sono un buon esempio di questo fenomeno. Nelle soluzioni ammoniacali la radiazione visibile e infrarossa è assorbita sia dagli elettroni solvatati sia dalla specie diamagnetiche. È significativo che lo spettro di assorbimento sia uguale per qualsiasi metallo alcalino presente in soluzione. Se la specie diamagnetica fosse un anione metallico, allora lo spettro differirebbe leggermente al variare del metallo. Evidentemente gli elettroni interagiscono fortemente solo con altri elettroni e non interagiscono fortemente con gli ioni metallici.

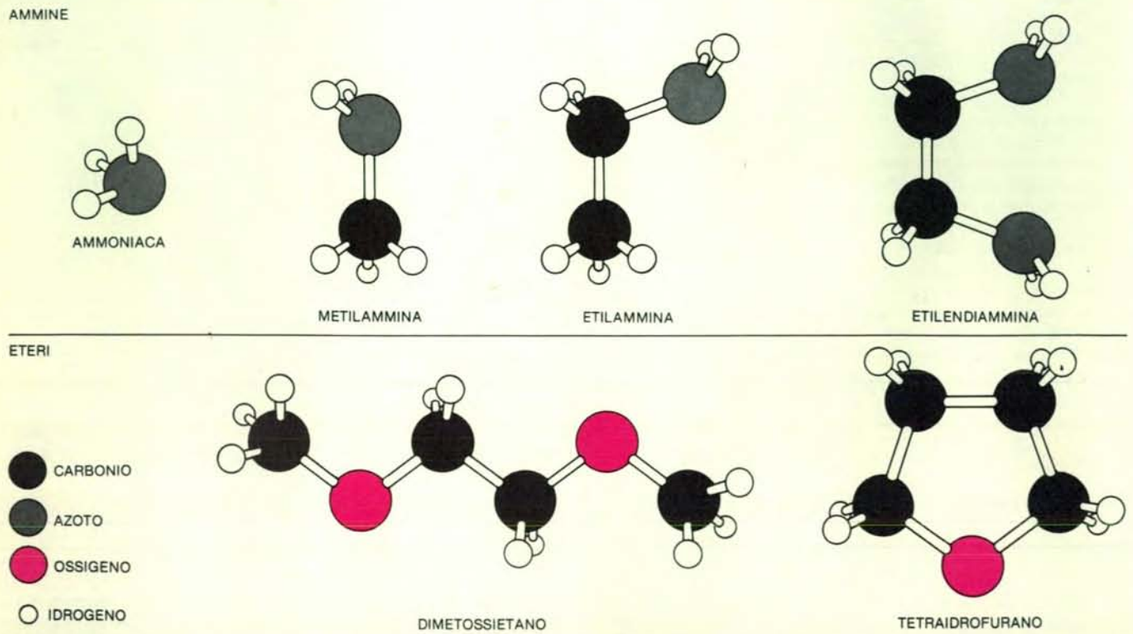
Potrebbe darsi che gli anioni metallici non esistano in ammoniaca semplicemente perché questa è un solvente troppo buono. Il catione potrebbe venire talmente stabilizzato dall'interazione con le molecole polari di ammoniaca da non rendere possibile l'esistenza di altre forme ioniche del metallo. Ogni anione eventualmente formato tenderebbe di conseguenza a scindersi in un catione e in due elettroni solvatati.

La scelta dei possibili solventi per i metalli alcalini è limitata. Molti liquidi non sono sufficientemente polari per sciogliere i metalli. Gli idrocarburi semplici, per esempio, come le frazioni del petrolio o il benzene, sono inerti nei confronti dei metalli alcalini. D'altro canto molti liquidi polari reagiscono con tali metalli formando dei composti. Si è già detto della reazione fra il sodio metallico e l'acqua, che è un solvente molto polare. Reazioni simili, fra cui talune molto violente, fanno eliminare intere classi di composti come gli alcoli, i chetoni e gli idrocarburi clorurati.

Una categoria di solventi dei metalli alcalini consiste di sostanze con molecole imparentate con l'ammoniaca. Per esempio la metilammina, che ha la formula CH_3NH_2, può venire considerata come una molecola di ammoniaca con un atomo di idrogeno sostituito dal gruppo metile $—CH_3$. Il litio si scioglie rapidamente in metilammina liquida, ma il sodio, il potassio, il rubidio e il cesio sono molto meno solubili. Un altro stretto parente dell'ammoniaca è l'etilendiammina, con la struttura $NH_2—CH_2—CH_2—NH_2$. Questa può venire considerata come un composto di condensazione di due molecole di metilammina. Nel mio laboratorio alla Michigan State University io e i miei collaboratori abbiamo trovato che l'etilendiammina è un solvente utile per i metalli alcalini.

Oltre alle ammine un altro gruppo importante di solventi dei metalli alcalini è costituito dagli eteri. Si tratta di una classe di composti organici contenenti il gruppo $—CH_2—O—CH_2—$. Il comune etere usato in anestesia è l'etere dietilico ($CH_3—CH_2—O—CH_2—CH_3$), che però non è un solvente dei metalli alcalini. Gli eteri più efficaci come solventi sono i polieteri, con più di un atomo di ossigeno, come il dimetossietano ($CH_3—O—CH_2—CH_2—O—CH_3$). Un altro solvente è il tetraidrofurano, un etere a struttura ciclica. Va detto che nessuna di queste sostanze è un buon solvente dei metalli alcalini e che esse non consentono la preparazione di soluzioni concentrate. In ogni caso consentono di sciogliere una quantità di metallo sufficiente per studiare le proprietà delle specie chimiche formate in soluzione.

Le proprietà ottiche di queste soluzioni sono significativamente diverse da quelle delle soluzioni ammoniacali. Si può ancora osservare la banda di assorbimento nel rosso e infrarosso associata alla presenza di elettroni liberi, ma a lunghezze d'onda minori si trovano delle caratteristiche di assorbimento nettamente diverse. Ciò che è più importante è che questa nuova banda di assorbimento si sposta al variare del metallo disciolto.



I solventi dei metalli alcalini appartengono a due categorie principali: ammoniaca e suoi derivati (ammine), che contengono azoto, e gli eteri, che contengono ossigeno. Questi liquidi sono sufficientemente polari per riuscire a sciogliere i metalli e, cosa ugualmente importante, non reagiscono con questi. Con l'eccezione dell'ammoniaca, si possono realizzare soltanto soluzioni con concentrazioni modeste di metallo.

Nel 1969 S. Matalon, Golden e Michael Ottolenghi suggerirono che questa nuova banda potesse venire attribuita agli anioni metallici. Questa proposta è risultata in seguito corretta.

È una cosa semplice dimostrare che una soluzione contiene una specie chimica con carica —1, ma l'anione metallico non è la sola specie che può avere questa proprietà. Per esempio potrebbe trattarsi di un catione metallico accompagnato da due elettroni solvatati debolmente legati. Un'altra possibilità è costituita dal fatto che due elettroni occupino un «orbitale espanso» avente al centro il catione, come nel vero anione, ma incorporante molecole di solvente. Oppure il catione potrebbe aver legati debolmente due elettroni strettamente legati fra di loro.

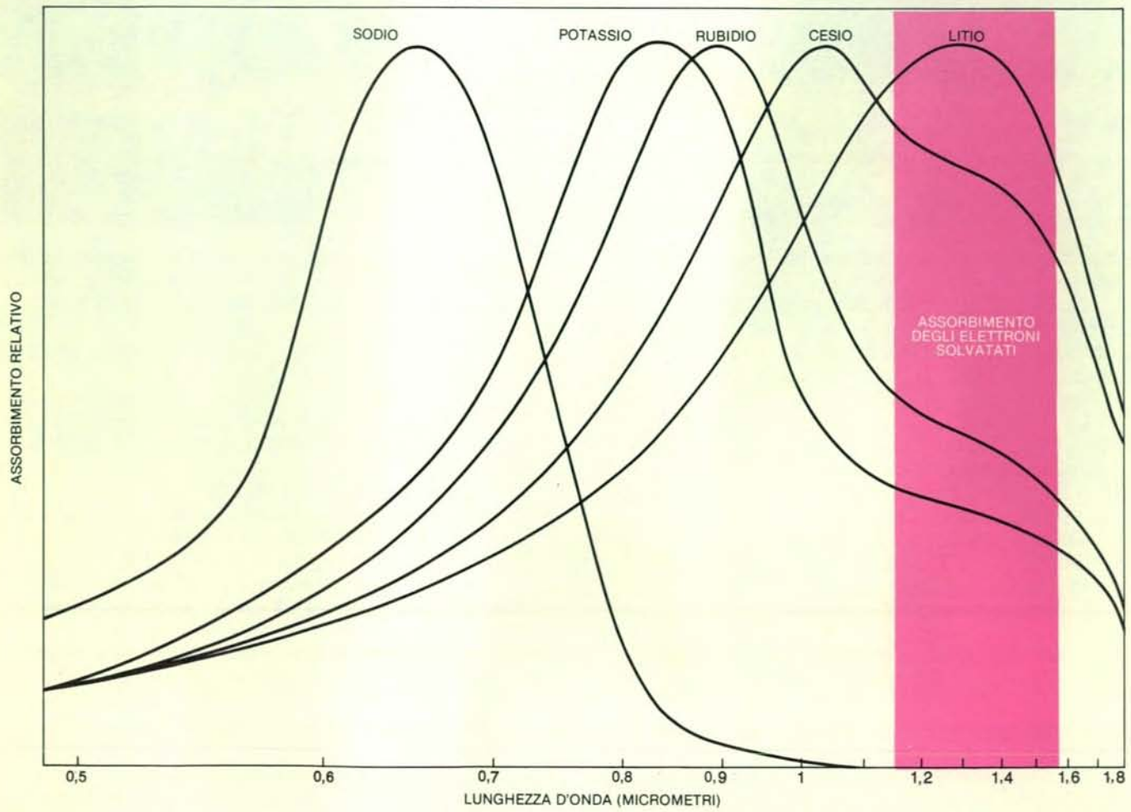
Tutte queste specie possono venire distinte in maniera non equivoca dal vero anione metallico. Tutte sono caratterizzate dalla presenza di un catione immerso nel solvente, le cui molecole di solvatazione devono avere l'orientamento tipico delle molecole attratte da uno ione positivo. I due elettroni stanno da qualche altra parte, fuori dello strato di solvatazione. Se invece si ha a che fare con un vero anione metallico i due elettroni sono legati strettamente allo ione e il solvente interagisce con questo come con un'entità singola, avente carica negativa.

La determinazione degli effetti ottici e magnetici nelle soluzioni dei metalli alcalini in ammine e in eteri fornisce prove convincenti della presenza di veri anioni metallici. Nel mio laboratorio si è dimostrato che la banda di assorbimento ottico è dovuta a due elettroni equivalenti e intercambiabili: è il caso dell'anione ma anche di qualche altra struttura. Insieme a Leon M. Dorfman della Ohio State University abbiamo misurato la velocità di formazione dell'anione Na^- a partire da elettroni solvatati e cationi Na^+. Altri ricercatori dimostrarono che la specie negativa poteva essere scissa in un catione e due elettroni solvatati, che successivamente si ricombinavano. Nel 1970 pareva molto probabile che l'esistenza dell'anione metallico fosse stata provata, ma non potevano venire ancora scartate altre interpretazioni. Uno degli ostacoli maggiori all'approfondimento degli studi era la scarsa solubilità dei metalli.

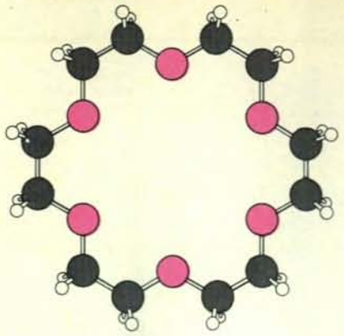
Nel 1970 Vincent A. Nicely, che lavorava nel mio laboratorio, suggerì che la solubilità dei metalli alcalini potesse venire aumentata con l'aiuto di una nuova classe di composti organici detti «eteri corona». Era già noto che questi eteri, sintetizzati da Charles J. Pedersen della E.I. du Pont de Nemours & Co., potevano aumentare la solubilità dei comuni sali dei metalli alcalini. Come suggerisce il loro stesso nome, gli eteri corona hanno la forma di un anello merlato. I cationi si annidano al centro dell'anello, che ha affinità per le cariche positive, e il complesso risultante può venire rapidamente solvatato da un liquido polare come un'ammina o un etere.

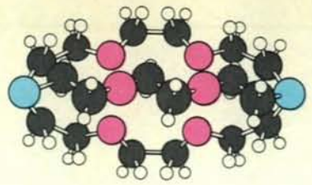
Dato che gli eteri corona non sono chimicamente molto diversi dai polieteri che avevamo lungamente impiegato come solventi, pareva ragionevole pensare che non avrebbero reagito in maniera violenta con i metalli. L'ipotesi risultò corretta e, come si sperava, la solubilità dei metalli aumentò di vari ordini di grandezza.

Un'altra categoria di sostanze che formano complessi con i metalli alcalini venne individuata più tardi da Jean-Ma-

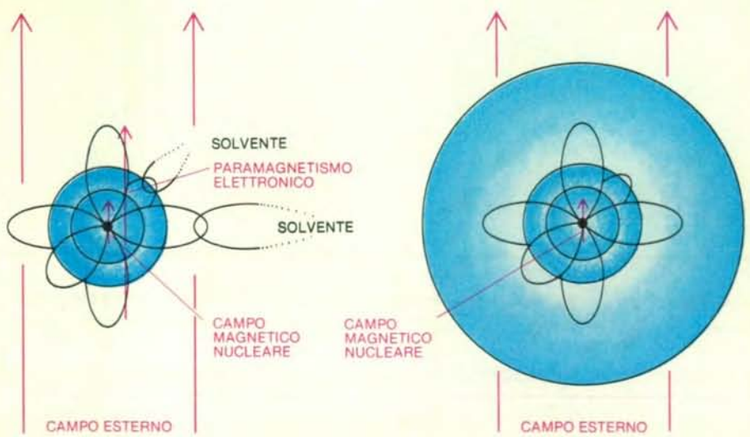


Lo spettro d'assorbimento delle soluzioni di metalli alcalini nelle ammine suggerisce la presenza di due specie assorbenti. Alla lunghezza d'onda di circa 1,4 micrometri, nella regione dell'infrarosso, il picco d'assorbimento è stato attribuito agli elettroni solvatati. Questo picco, a lunghezze d'onda inferiori, è ora attribuito agli anioni dei metalli alcalini. È significativo che cada a una lunghezza d'onda diversa per i vari metalli. Questi due picchi si possono osservare negli spettri del potassio, del rubidio e del cesio. Nello spettro del litio si osserva solo il picco associabile agli elettroni solvatati e non sembra che vi siano in soluzione degli anioni metallici. Al contrario il sodio pare formare anioni con tale facilità che non si osserva la presenza di elettroni solvatati. Gli spettri si riferiscono a soluzioni in etilendiammina.





Gli eteri corona e i criptanti imprigionano i cationi dei metalli alcalini migliorandone la solubilità e incidentalmente favorendo la formazione degli anioni metallici. In alto sono mostrate schematicamente le strutture dei due tipi di molecole. I criptanti sono più efficienti degli eteri corona perché la loro nicchia tridimensionale isola più efficacemente lo ione dal solvente. La rappresentazione tridimensionale del criptante (*in basso*) mostra come un catione sodio si adatti bene alla cavità della molecola. Nelle soluzioni e nei sali i criptanti stabilizzano i cationi dei metalli alcalini, impedendo loro di ricombinarsi con gli ioni formando il metallo.



Le proprietà magnetiche di un atomo sono influenzate dal suo stato di ionizzazione e dall'ambiente circostante. Il campo magnetico del nucleo può venire misurato con la tecnica della spettroscopia di risonanza magnetica nucleare e tale misura fornisce informazioni indirette sullo stato degli elettroni atomici. Negli atomi, o ioni, che hanno i loro elettroni più esterni in orbitali *p* (*a sinistra*) le interazioni con le molecole del solvente tendono ad aumentare l'effetto che un campo magnetico esterno ha sul campo del nucleo. Questo effetto è detto paramagnetismo. Nel caso che gli elettroni esterni si trovino in orbitali *s* (*a destra*) l'effetto paramagnetico è assente.

rie Lehn dell'Università Louis Pasteur di Strasburgo. Per analogia con il classico nome di legante, che definisce le sostanze che si legano ai metalli formando dei complessi, Lehn chiamò i suoi composti «criptanti», in quanto la loro struttura spaziale forma una nicchia, o cripta, entro la quale viene racchiuso il catione. Esso viene così sequestrato e schermato dalle molecole circostanti di solvente. I criptanti di nostro interesse hanno tutti due atomi di azoto connessi da tre catene organiche abbastanza lunghe. Ogni catena è un etere con uno, due o tre atomi di ossigeno. La semplice designazione del numero degli atomi d'ossigeno nella catena è sufficiente per designare un determinante criptante Lo 1,1,1-criptante, con un solo atomo di ossigeno per catena possiede una piccola cavità che può racchiudere solo un idrogenione oppure un catione litio Li^+. Le cavità più grandi dei criptanti 2,2,1- e 2,2,2- sono ideali per sequestrare lo ione Na^+.

I criptanti possono fornire enormi aumenti della solubilità. Per esempio, in etilammina pura il sodio è praticamente insolubile, ma in presenza di 2,2,2-criptante si possono sciogliere diversi grammi per litro di solvente con un aumento di 400 000 volte nella solubilità.

È la stabilità del complesso metallo-criptante che rende possibile tali elevate solubilità. Paradossalmente la stabilità del catione entro la cavità del criptante rende possibile il raggiungimento di elevate concentrazioni di anioni metallici. In effetti aggiustando le concentrazioni del metallo e del criptante è possibile controllare le specie ioniche presenti in soluzione.

Quando gli atomi del metallo e le molecole di criptante sono presenti in numero uguale, praticamente tutti i cationi vengono catturati entro le nicchie della molecola organica. Questa regola vale per il potassio, il rubidio e il cesio. Ne consegue che in queste soluzioni la specie negativa predominante è l'elettrone solvatato, dal momento che semplicemente non vi sono atomi disponibili per fare l'anione. Se il metallo alcalino è presente in eccesso, il catione intrappolato rimane la specie carica positivamente predominante, ma ora gli atomi metallici in eccesso possono combinarsi con gli elettroni per formare anioni dei metalli alcalini. Queste regole generali valgono per tutti i metalli alcalini con l'eccezione del litio e del sodio (il francio non è stato studiato). Nelle ammine e negli eteri il litio pare riluttante a formare anioni in qualsiasi condizione. Il sodio, d'altro canto, forma un anione così stabile che gli elettroni solvatati si osservano raramente.

Nelle soluzioni concentrate che si possono preparare con l'aiuto degli eteri corona e dei criptanti è molto più facile studiare le proprietà degli anioni dei metalli alcalini. Per esempio diviene possibile impiegare la spettroscopia di risonanza magnetica nucleare. Si tratta di una tecnica che consente di selezionare un elemento particolare presente in solu-

zione per studiarne il suo stato di legame.

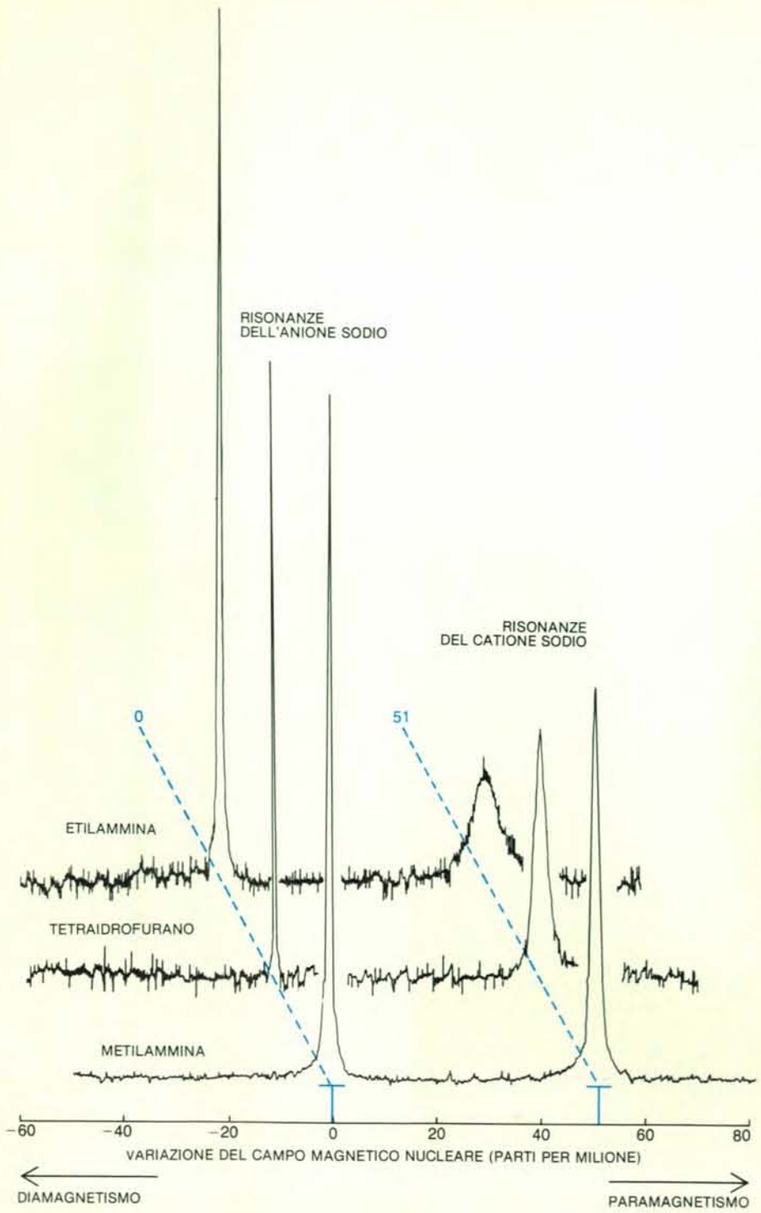
I nuclei atomici possiedono uno spin esattamente come gli elettroni e quindi anche i nuclei possono generare dei campi magnetici. Tutti i nuclei dei metalli alcalini sono magnetici: il nucleo del sodio, per esempio, possiede quattro stati di spin (in confronto ai due dell'elettrone) e quindi quattro possibili orientamenti del suo campo magnetico rispetto a un campo esterno. Il magnetismo nucleare non ha nessuna importanza per quanto riguarda la chimica dell'elemento ma è un utile rivelatore del campo magnetico elettronico, il quale invece è dipendente dallo stato di legame e quindi dal comportamento chimico specifico dell'atomo.

In pratica il campo magnetico nucleare viene studiato forzando il nucleo a mutare il suo stato di spin. Dal momento che ogni stato di spin richiede un diverso orientamento del campo magnetico nucleare rispetto a un campo esterno, ognuno di questi stati possiede un diverso livello energetico caratteristico. La differenza fra i livelli energetici è proporzionale all'intensità del campo esterno percepito dal nucleo. L'energia necessaria per far passare il nucleo al livello energetico superiore può venire fornita da un fotone, cioè da un quanto di energia elettromagnetica. Allo stesso modo il nucleo può ricadere al livello inferiore emettendo un fotone. L'energia dei fotoni in gioco è in genere equivalente a quella delle radiofrequenze comprese fra circa 2 e 300 megahertz. Quando sia il campo esterno sia il valore della radiofrequenza sono aggiustati opportunamente, il nucleo entra in risonanza e la radiazione viene fortemente assorbita.

Il ricercatore che studia la risonanza magnetica di nuclei in soluzione può controllare il campo esterno nel quale è immerso il campione, ma questo non è necessariamente il campo a livello dei nuclei. Gli elettroni che circondano i nuclei possono infatti creare un campo opposto al campo esterno (comportamento diamagnetico) oppure tale da sommarsi a questo (comportamento paramagnetico). Per compensare la schermatura diamagnetica degli elettroni il campo esterno deve venire aumentato, mentre il comportamento paramegnetico richiede una diminuzione del campo. Ne segue che questi «spostamenti chimici» contengono delle informazioni riguardanti i dintorni del nucleo.

È proprio la tecnica della risonanza magnetica nucleare che ha fornito prove convincenti dell'esistenza di anioni dei metalli alcalini in soluzione. È possibile comprendere come la risonanza magnetica consenta di distinguere fra anioni veri e propri e altre specie possibili esaminando lo stato degli elettroni dell'atomo di sodio in diverse situazioni.

Ai fini della spettroscopia di risonanza magnetica nucleare uno ione Na^+ può venire considerato un riferimento di zero del campo esterno, avendo la configurazione elettronica non magnetica dei gas nobili. La situazione è però diversa per un catione Na^+ in soluzione. Dobbiamo ricordare che gli elettroni più esterni del catione si trovano in orbitali *p* e possiedono quindi oltre allo spin anche un momento angolare orbitale. Quando il catione è isolato i momenti dei sei elettroni *p* si annullano esattamente a due a due, ma l'interazione con le molecole di solvente disturba il bilancio. Il risultato di tale interazione è che gli orbitali *p* acquistano un movimento di precessione intorno al nucleo, dando quindi origine a un campo magnetico che si somma a quello esterno. In altre parole il catione in soluzione mostra uno spostamento chimico di tipo paramagnetico. Lo stesso vale per un catione sequestrato entro una molecola di criptante. L'entità dello spostamento dipende principalmente dalla natura delle molecole a contatto diretto col catione.



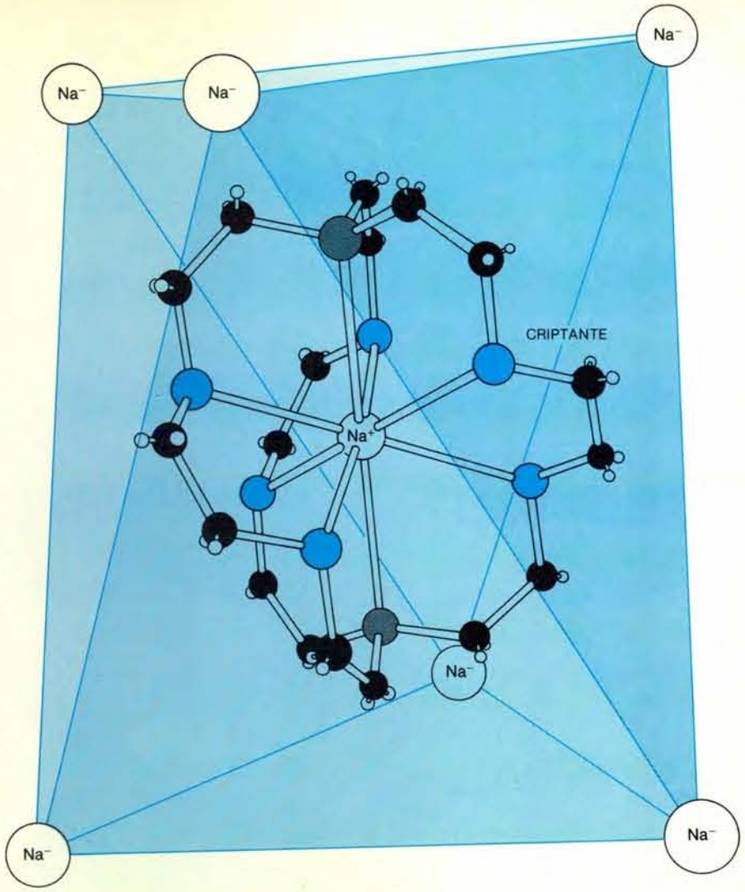
Gli spettri di risonanza magnetica nucleare forniscono una prova convincente dell'esistenza di veri anioni metallici in varie soluzioni di sodio. Le soluzioni concentrate necessarie per le misure sono state ottenute con l'aiuto di criptanti. Lo zero della scala è scelto in modo da corrispondere al campo magnetico indisturbato del nucleo di sodio nel catione gassoso. Le deviazioni apparenti da questo valore nel campo nucleare misurato sono dovute ai campi elettronici che rinforzano (paramagnetismo) o diminuiscono (diamagnetismo) il campo nucleare. I picchi larghi e spostati in senso paramagnetico di 51 parti per milione sono associati ai cationi imprigionati nel criptante. I picchi più stretti non spostati sono le risonanze degli anioni sodio.

Un anione di un metallo alcalino possiede solamente orbitali completi e, qualora fosse isolato, dovrebbe avere uno spettro di risonanza magnetica nucleare praticamente identico a quello del catione isolato considerato come riferimento. Ma quel che è più importante è che lo spettro non dovrebbe mostrare nessuno spostamento chimico in soluzione. Infatti gli elettroni più esterni dell'anione occupano un orbitale *s* e non possiedono alcun momento angolare orbitale. Le loro interazioni col solvente non danno origine ad alcun campo magnetico.

Importanti studi preliminari sulla risonanza magnetica nucleare dei cationi del sodio complessati entro criptanti vennero eseguiti nel mio laboratorio da Joseph M. Ceraso nel 1973. Egli dimostrò che i cationi Na^+ rimangono entro la cavità del criptante abbastanza a lungo per poter essere distinti da quelli liberi in soluzione. Successivamente l'attenzione venne rivolta verso le soluzioni contenenti sia il catione legato al criptante sia la specie ionica sospettata di essere l'anione Na^-. Si poterono osservare due segnali. Uno possedeva il tipico spostamento paramagnetico del catione legato. L'altro non presentava nessuno spostamento rispetto alla posizione calcolata per l'anione sodio isolato. Solo in un vero anione gli elettroni *s* possono schermare efficacemente il nucleo. Altre specie ioniche possono ben avere una carica negativa ma avrebbero presentato la firma magnetica del catione in soluzione. Risultati simili sono stati ottenuti in soluzioni di rubidio e di cesio.

Se un metallo alcalino può comportarsi in soluzione come un anione, dovrebbe giocare il medesimo ruolo in un sale precipitato da tale soluzione. È possibile immaginare, per esempio, un analogo ipotetico di un sale semplice come il cloruro di sodio, ma contenente sodio sia sotto forma di catione che di anione. I calcoli indicano che un tale sale sarebbe meno stabile del sodio metallico, che quindi si formerebbe spontaneamente da esso, però la differenza di stabilità fra sale e metallo è piccola. Noi abbiamo preparato dei sali più complicati, nei quali il catione è un complesso fra sodio e criptante. Il criptante aumenta la stabilità del sale quanto basta per evitare la formazione del metallo.

Mei Tak Lok e Frederick J. Tehan tentarono nel mio laboratorio di preparare un sale da un'adatta soluzione di sodio e criptante in etilammina. Dato che si sospettava che il sale avrebbe avuto la tendenza a scindersi per formare sodio metallico, si evaporò il solvente rapidamente, nella speranza di ottenere un precipitato prima che i cationi avessero il tempo di sfuggire dal criptante. I risultati furono incoraggianti: ottenemmo infatti un film dorato che, all'aspetto, era completamente differente tanto dal sodio metallico quanto dal criptante.



La struttura dei cristalli precipitati da soluzioni sodiche si risolve in strati alternati di anioni sodio e del complesso formato dal criptante e dal catione Na^+. Sotto alcuni aspetti fondamentali questi cristalli rassomigliano fortemente a quelli dei sali più convenzionali, come il cloruro di sodio. Per esempio il numero degli anioni e dei complessi catione-criptante deve essere lo stesso. L'elemento che distingue i due tipi di cristallo è la reattività degli anioni metallici.

In pratica la preparazione del sale risultò più facile del previsto e l'evaporazione rapida del solvente era del tutto inutile. Ceraso, nel corso della preparazione di soluzioni di sodio e criptante per misure di risonanza magnetica nucleare, aveva già notato la formazione spontanea di un precipitato dorato che avveniva quando la soluzione veniva raffreddata con ghiaccio secco a —78 °C. Più tardi Lok e Tehan trovarono che il materiale ha una struttura cristallina corrispondente a un sale con la formula Na^+2,2,2-criptante · Na^-.

A causa del colore dorato dei cristalli pensammo in un primo momento che il composto si sarebbe comportato come un metallo, che è una sostanza caratterizzata dalla presenza di un gran numero di elettroni mobili. Risultò invece che si trattava di un semiconduttore, una sostanza cioè in cui gli elettroni di valenza sono praticamente tutti legati, anche se debolmente, a un atomo determinato.

In soluzione di etilammina i cristalli crescono sotto forma di sottili lamelle esagonali. Tehan e B.L. Barnett hanno determinato la struttura dei cristalli, confermando l'ipotesi che si trattasse di un sale. Strati del complesso fra catione e criptante si alternano a strati dell'anione metallico. Gli anioni Na^- rassomigliano agli ioni ioduro per dimensioni e posizione occupata nel reticolo cristallino.

Il nuovo composto allo stato solido è estremamente reattivo nei confronti dell'aria, dell'umidità e in generale di qualsiasi accettore di elettroni. Anche quando è conservato nel vuoto non presenta una stabilità indefinita se non a basse temperature. A temperatura ambiente i cristalli lentamente scuriscono e perdono la loro lucentezza metallica man mano che gli anioni sodio attaccano le molecole di criptante. Ciononostante, lavorando rapidamente, siamo riusciti a determinare il punto di fusione dei cristalli. Esso è a 83 °C, 15 gradi sopra il punto di fusione del criptante puro. La fusione

si accompagna alla decompressione in sodio metallico e criptante libero.

Recentemente io, Michael G. DaGue e Michael R. Yemen siamo riusciti a spiegare le ragioni del sorprendente aspetto metallico di questo composto non metallico. La lucentezza metallica è un'impressione soggettiva dovuta a uno spettro d'assorbimento della luce con una banda rapidamente crescente verso le lunghezze d'onda maggiori. Nei metalli uno spettro d'assorbimento molto simile è dovuto agli elettroni liberi entro il reticolo cristallino. I cristalli contenenti l'anione sodio possiedono per caso uno spettro di quel tipo anche se non possiedono elettroni liberi in gran quantità. L'aspetto metallico è quindi casuale. Anche la struttura cristallina del sale, con grandi facce come un metallo, contribuisce all'illusione.

Dopo aver identificato il sale dell'anione sodio siamo riusciti a preparare composti contenenti potassio, rubidio e cesio che paiono avere una struttura analoga. L'identificazione definitiva di questi composti è stata ostacolata dalla loro comparativamente elevata instabilità. A temperatura ambiente il sale sodico ha un tempo di semidecomposizione di vari giorni, il corrispondente composto potassio si decompone a metà in meno di un minuto. I sali dei metalli più pesanti devono quindi essere mantenuti a temperature inferiori ai —40 °C, dove sembrano rimanere stabili per lunghi periodi.

Dato che le soluzioni di anioni dei metalli alcalini e dei corrispondenti cationi intrappolati danno origine a composti solidi, sembra ragionevole pensare alla possibilità di ottenere sostanze cristalline simili a partire da cationi intrappolati ed elettroni solvatati. Una possibilità è data da una sorta di «metallo espanso», nel quale gli elettroni sono liberi di muoversi non in un reticolo costituito da cationi semplici bensì da un complesso metallo-criptante. I tentativi di preparare un composto di questo tipo sono finora falliti. Peraltro, evaporando rapidamente il solvente sono stati ottenuti dei film che paiono avere una struttura diversa. Questi film hanno un colore blu intenso e sono fortemente paramagnetici, cosa che fa pensare che gli elettroni siano confinati in prossimità di un particolare catione invece che essere liberi di muoversi nel reticolo. Misure preliminari sembrano indicare che il solido contiene esattamente un elettrone spaiato per ogni catione. Se questo risultato verrà confermato allora i film non sono per nulla metallici ma sono invece il primo esempio di una nuova classe di sali che possiamo chiamare «elettruri».

Si può ora ritenere provato che tutti i metalli alcalini, con l'eccezione del litio, formano veri anioni. I calcoli però mostrano che anche l'anione litio dovrebbe essere stabile in condizioni opportune e pertanto stiamo continuando le ricerche in tale direzione. Con ogni probabilità abbiamo scoperto solo un numero limitato di composti contenenti anioni dei metalli alcalini. Dato che le dimensioni del criptante possono essere variate facilmente, è possibile che vari cationi possano venire combinati con un anione qualsiasi. Una possibilità ovvia è quella di sali contenenti il catione intrappolato di un metallo alcalino e l'anione di un altro. Più in generale il criptante può inglobare anche cationi diversi da quelli dei metalli alcalini. Per esempio si può pensare di ottenere sali dei metalli alcalino-terrosi, come il bario, il cui catione ha carica +2, oppure delle terre rare, come il lantanio, che hanno cationi con carica +3. I sali di questi cationi dovrebbero necessariamente contenere una concentrazione elevata degli anioni dei metalli alcalini, due o tre per ogni catione. E infine i criptanti potrebbero anche venire eliminati. Certi cationi organici, fra i quali, per esempio, certi derivati organici dello ione ammonio, potrebbero sopravvivere senza protezione in presenza degli anioni dei metalli alcalini e forse anche di elettroni solvatati.

Ora col tuo denaro puoi fare molto di piú.

rivolgersi alle banche che espongono questo simbolo.

L'occhio composto degli insetti

Possiede centinaia di faccette, ciascuna delle quali corrisponde a un ommatidio. Le dimensioni, la struttura e la direzione dell'asse ottico degli ommatidi dipendono dalle necessità insite nel modo di vita degli insetti

di G. Adrian Horridge

Esiste un numero sterminato di insetti; alcuni sono attivi solo nella piena luce solare, altri sia al sole, sia all'ombra, altri solo al crepuscolo. Alcuni, come la mosca domestica, volano ad alta velocità e compiono rapide giravolte. Altri, come la libellula, volano in linea retta, sono in grado di compiere manovre e di librarsi. Altri infine, come la mantide religiosa, rimangono immobili per ore intere. La possibilità di vita di tutte queste creature si basa sulla vista e sulla percezione dell'ambiente esterno mediante occhi composti, caratteristici per la loro abbondante sfaccettatura. Le abitudini e le necessità visive di ciascun insetto, però, sono molto diverse.

Come funziona il meccanismo di raccolta delle immagini d'un occhio composto? Fino a che punto rivela le funzioni per le quali l'insetto usa i propri occhi? Fino a che punto la piccolezza delle faccette dell'occhio ne limita la sensibilità? Qual è la soglia di sensibilità dell'occhio a basse intensità luminose con le piccole lenti di cui sono dotati gli ommatidi? Come riesce l'occhio composto degli insetti ad arrivare a un compromesso tra la risoluzione ottica e la sensibilità necessaria per superare i segnali ottici di disturbo o rumore?

Un occhio composto è costituito da ommatidi, ossia da minuscoli occhi che puntano in direzioni diverse. Ogni ommatidio è, a sua volta, costituito da una lente che focalizza la luce su parecchie cellule recettrici che partecipano alla formazione di un organo fotosensibile comune, chiamato rabdoma. In moltissimi insetti ogni ommatidio è un singolo elemento di raccolta con il proprio asse ottico; qui saranno presi in esame solo occhi d'insetto con questo tipo di ommatidio. La luce assorbita del rabdoma produce un segnale che viene trasmesso ai centri nervosi attraverso un nervo ottico. Un'altra domanda che sorge a questo punto è: come possiamo determinare la direzione in cui gli ommatidi stanno guardando? E che cosa ci può dare informazioni sul modo in cui l'insetto percepisce il mondo esterno?

Guardando l'occhio d'un insetto, spesso vediamo una macchia nera nel centro dell'occhio stesso. Quando l'insetto ruota il capo, la macchia nera punta sempre nella direzione dell'osservatore. Tale macchia prende il nome di pseudopupilla: le faccette che vi sono contenute sembrano nere perché riflettono meno luce nella direzione dell'osservatore rispetto a quelle situate nel resto dell'occhio. In altri termini, gli ommatidi nel centro della pseudopupilla guardano direttamente verso l'osservatore.

Per disegnare una mappa degli assi ottici di ciascun ommatidio, si monta il capo dell'insetto (oppure l'animale intero) al centro d'un goniometro, strumento che misura gli angoli. Si misura l'angolo necessario per spostare il centro della pseudopupilla a una distanza di cinque faccette e la misurazione viene ripetuta a intervalli di cinque faccette fino a disegnare la mappa di tutto l'occhio. I medesimi dati si possono registrare più facilmente fotografando l'occhio a intervalli di 5 o 10 gradi. In questo caso le faccette sono contrassegnate da particelle di polvere e le posizioni della pseudopupilla rispetto allo schema delle faccette contrassegnate sono misurate servendosi delle fotografie. La posizione dell'asse di ogni quinta faccetta viene poi registrata su un reticolo a coordinate angolari. Lo schema risultante è una mappa del modo in cui l'occhio composto prende in esame gli elementi vivi presenti nell'ambiente che lo circonda. Finora sono state eseguite solo poche mappe parziali, come quella preparata per la mosca domestica dai biofisici olandesi D.G.M. Beersma, Doekele G. Stavenga e Jan W. Kuiper. Solo strisce molto strette possono essere rappresentate su una mappa senza incontrare le ben note difficoltà legate allo sviluppo di una superficie sferica in un foglio piano.

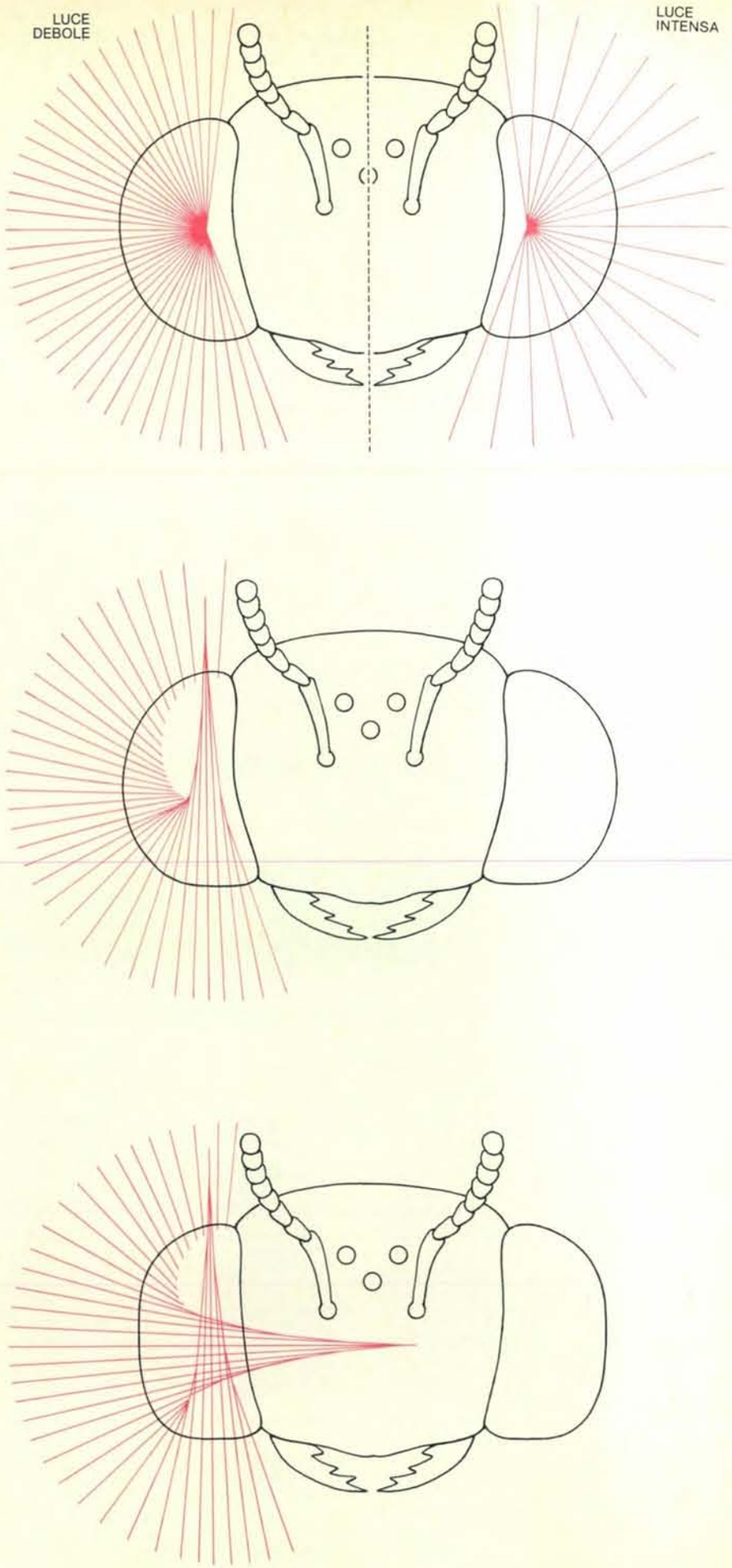
Per capire la funzione dell'occhio d'un insetto è importante definire due tipi di angolo. Il primo è l'angolo tra gli assi ottici di due ommatidi adiacenti: tale angolo, denominato angolo interommatidiale e indicato con $\Delta\phi$, determina la «compattezza» di rilevamento dei dati del campo visivo da parte dell'occhio composto. Il secondo è l'angolo che definisce il campo visivo d'un singolo ommatidio e viene indicato con $\Delta\rho$. La forma del campo equivale in pratica alla sensibilità dell'ommatidio rispetto a una sorgente luminosa puntiforme espressa come distanza angolare della sorgente dall'asse ottico. La sensibilità dell'ommatidio è massima lungo l'asse ottico e diminuisce proporzionalmente alla distanza angolare dell'asse. L'ampiezza del campo è definita come l'angolo sotteso dal diametro del campo nel punto in cui la sensibilità è diminuita del 50 per cento.

Le mappe dell'occhio rivelano alcune sorprese circa il modo in cui gli insetti vedono l'ambiente che li circonda. Innanzitutto le mappe dimostrano che il modello strutturale degli assi ottici non è uguale al modello strutturale delle faccette esagonali disposte sulla superficie dell'occhio. Faccette che si trovano a distanze uguali sulla superficie dell'occhio non possiedono necessariamente assi ottici disposti secondo angoli interommatidiali uguali. Infatti in alcune regioni dell'occhio gli assi ottici degli ommatidi possono essere inclinati l'uno rispetto all'altro in modo da essere quasi paralleli, fornendo così per una porzione dell'ambiente visivo un numero di informazioni visive superiore al normale. In quella regione la pseudopupilla è di solito più grande rispetto al resto dell'occhio, poiché molte faccette sono rivolte verso quella direzione.

Tale area dell'occhio composto degli insetti ha la medesima funzione della *fovea centralis* della retina dei vertebrati (e anche dell'uomo). Nei vertebrati la fovea è un'area in cui le cellule fotorecettrici sono situate a minor distanza le une dalle altre e perciò forniscono una immagine ad alta risoluzione. In effetti negli occhi degli animali diurni (tra cui l'uomo) i recettori debbono essere molto stretti per essere così vicini gli uni agli altri e la sensibilità di ciascun recettore è di conseguenza ridotta. Quindi nei vertebrati la fovea rappresenta un adattamento che permette una visione ottimale con luce intensa.

Negli insetti, comunque, le faccette della fovea sono di solito più grandi di quelle del resto dell'occhio. Poiché l'angolo interommatidiale nella fovea risulta ridotto, la ragione che comprende la fovea è otticamente più appiattita rispetto al resto dell'occhio: in altre parole, ha un raggio di curvatura maggiore. Perciò, per far posto a una fovea in una regione d'un occhio composto, l'occhio deve possedere un raggio ottico di curvatura minore nelle altre zone. Quindi alcune aree dell'occhio risultano impoverite per la raccolta dai dati visivi a favore di un'altra area che a sua volta può essere arricchita.

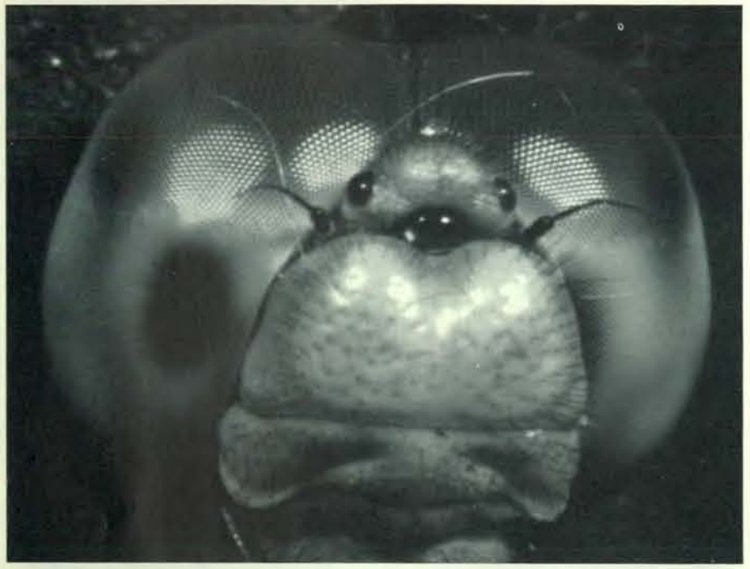
Le faccette di dimensioni maggiori permettono agli ommatidi corrispondenti un potere di risoluzione maggiore. Per qualsiasi lente la capacità di risolvere i particolari più fini dipende da due fattori: l'apertura, o diametro, della lente (D) e la lunghezza d'onda della luce (λ) focalizzata per mezzo della lente. (A causa della difficoltà di misurare l'apertura di una lente esagonale, D viene calcolata come distanza tra i centri di due faccette adiacenti). La luce proveniente da una sorgente puntiforme lontana, passando attraverso la lente, non viene focalizzata in un punto ma risulta sfuocata, diffusa su un cerchio nel piano focale della lente. L'intensità della luce in questo cerchio sfuocato (o cerchio di confusione) è maggiore nel centro e decresce verso la periferia. È opportuno definire l'ampiezza del cerchio di confusione come il diametro in cui l'intensità della luce è diminuita del 50 per cento: tale ampiezza viene indicata con Δa. Dalla teoria della diffrazione è noto che l'ampiezza angolare del cerchio di confusione in radianti è uguale alla lunghezza d'onda della luce



L'occhio composto degli insetti è costituito da ommatidi. Le dimensioni effettive di qualsiasi parte di questo tipo di occhio sono stabilite dal suo raggio ottico, che è definito come raggio misurato dal punto in cui gli assi ottici di ommatidi adiacenti si intersecano. Qui possiamo osservare diversi tipi di occhio composto, visti da un punto situato sopra il dorso dell'insetto. Se l'occhio dell'insetto fosse un emisfero e gli ommatidi di uguale apertura guardassero perpendicolarmente alla superficie dell'occhio, tutto l'occhio avrebbe un solo raggio ottico, e raccoglierebbe dati sull'ambiente visivo in tutte le direzioni con uguale densità (*in alto*). La sensibilità di un occhio composto dipende in parte dalla dimensioni delle sue faccette e perciò dalle sue dimensioni generali. Tuttavia, in seguito a misurazioni compiute su occhi composti realmente esistenti, si è trovato che in molti il raggio ottico nella regione frontale è maggiore rispetto alle altre regioni (*al centro*). In questo tipo d'occhio la porzione di ambiente che sta davanti viene esaminata in modo più accurato rispetto alle altre zone. Quest'area dell'occhio degli insetti viene chiamata fovea, per analogia con la fovea dell'occhio dei vertebrati. Sulla fovea le faccette sono di solito più grandi, il che è reso possibile dal fatto che qui l'occhio ha un raggio ottico di maggiori dimensioni. Alcuni insetti (*in basso*) possiedono una seconda fovea che guarda verso l'alto o lateralmente: le grosse libellule possiedono fovee di tutti questi tipi.

divisa per il diametro della lente ($\Delta a = \lambda/D$). (Un radiante corrisponde a un angolo giro diviso per 2π, cioè a circa 57 gradi). Quindi, per ogni lunghezza d'onda, più la lente è ampia, minore è il cerchio di confusione di ciascuna sorgente puntiforme e più nitida è l'immagine.

Negli insetti per i quali il possedere una vista acuta è fondamentale, la efficienza della lente di ciascun ommatidio è eccellente e la parte di recettore che capta la luce (in realtà la punta del rabdoma) giace proprio nel piano focale della lente. Esistono due limiti fondamentali alla risoluzione da parte del singolo ommatidio. Il primo è l'ampiezza del cerchio di confusione: una faccetta di grandi dimensioni sull'occhio di un insetto può presentare un potere di risoluzione superiore rispetto a una piccola, poiché concentra la luce in un cerchio di confusione più piccolo. Il secondo limite riguarda il diametro minimo compatibile con la funzionalità del recettore; si è trovato che il valore in oggetto è di circa 0,5 micrometri.



Sull'occhio della libellula *Diplacodes nematodes* appare la cosiddetta pseudopupilla, che nelle fotografie si distingue come una macchia nera. Quando il capo dell'insetto ruota, anche la pseudopupilla si sposta, in modo da puntare sempre verso l'osservatore. La pseudopupilla è nera, perché le faccette di questa zona riflettono meno luce nella direzione dell'osservatore rispetto alle faccette delle altre zone dell'occhio. Le faccette situate nel centro della pseudopupilla guardano direttamente verso l'osservatore. In generale, dove la pseudopupilla è più grande (*in alto*) vi sono parecchi ommatidi che stanno guardando in quella direzione. Nella parte laterale dell'occhio (*in basso*) la pseudopupilla è allungata in senso verticale, perché gli assi ottici nelle file verticali formano tra loro angoli pari a circa la metà di quelli fra gli assi degli ommatidi delle file orizzontali.

Se una sorgente luminosa puntiforme si trova in corrispondenza dell'asse ottico dell'ommatidio, la porzione più brillante del cerchio di confusione risulterà centrata sulla punta del rabdoma. Se la sorgente puntiforme si sposta attraverso il campo visivo, esce dall'asse ottico, e il massimo d'intensità luminosa del cerchio di confusione esce dal rabdoma. Perciò il rabdoma intercetterà una frazione progressivamente più piccola e meno intensa del cerchio di confusione. Se il rabdoma è relativamente stretto, esso può misurare con precisione la variazione d'intensità luminosa attraverso il cerchio di confusione e il suo potere risolutivo viene limitato solamente dalla diffrazione della luce che ha dato origine al cerchio di confusione.

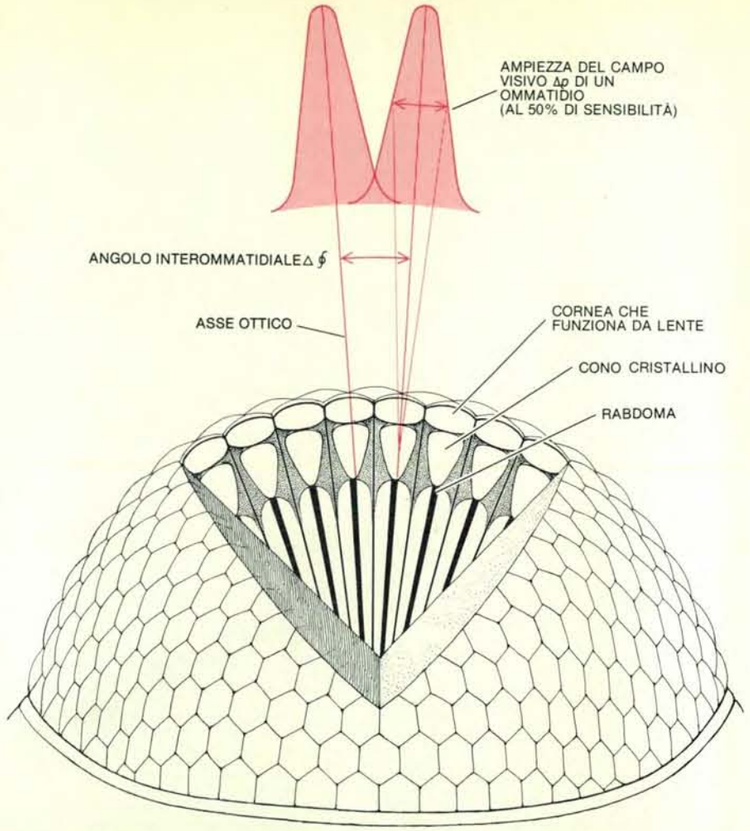
Per la stessa ragione, comunque, un rabdoma stretto capterà solo una piccola frazione della luce totale nel cerchio di confusione e la sua sensibilità sarà bassa. Se il rabdoma è più largo, riuscirà a captare una quantità maggiore di luce nel cerchio di confusione, ma il suo potere risolutivo sarà inferiore perché non potrà nel contempo definire i particolari. Quindi la sensibilità angolare ottimale del rabdoma è funzione sia dell'ampiezza del cerchio di confusione, Δa, sia del diametro d del rabdoma. In ogni occhio esiste un compromesso tra il potere di risoluzione, limitato unicamente dalla diffrazione e dalla ampiezza minima del rabdoma, e la sensibilità che, come vedremo, è importantissima negli occhi dotati di lenti piccole come quelle di cui stiamo trattando.

Questo compromesso può essere spiegato con un'analogia: il modo in cui il profilo del getto d'acqua di una doccia può essere misurato da contenitori di dimensioni diverse. Il diametro del getto corrisponde all'ampiezza del cerchio di confusione e le gocce d'acqua corrispondono ai fotoni. Una piccola scodella spostata attraverso il getto potrà misurare rapidamente, o risolvere, il profilo d'intensità del getto, o distinguere rapidamente i confini di parecchi spruzzi vicini gli uni agli altri. Ma in ogni punto la scodella può intercettare solo una piccola porzione del getto, cioè la sua sensibilità è bassa. Il getto viene intercettato molto più facilmente da una grossa tinozza: la tinozza è perciò più «sensibile» d'una scodella, ma pur raccogliendo tanta acqua non riesce a distinguere i particolari, a distinguere, cioè, gli spruzzi vicini. Tuttavia quando il flusso dell'acqua è lento, solo i getti di maggiore ampiezza possono essere svelati, e la tinozza fornisce il mezzo migliore per misurare la distribuzione del getto.

Se il diametro d d'un rabdoma circolare è ridotto in confronto all'ampiezza Δa del cerchio di confusione, il campo visivo $\Delta\varrho$ del rabdoma è uguale all'ampiezza angolare del cerchio di confusione. Se il rabdoma è abbastanza largo per riuscire a captare il 95 per cento di tutta la luce del cerchio di confusione, il campo visivo $\Delta\varrho$ del rabdoma risulta essere doppio rispetto all'ampiezza angolare del cerchio di confusione con una intensità del 50 per cento ossia $2\,\Delta a$. L'aumento del diametro del rabdoma oltre questo valore ha come risultato una continua perdita di risoluzione senza un corrispondente aumento di sensibilità a una sorgente luminosa puntiforme. Nell'analogia della doccia, il fatto di avere una tinozza più grande del diametro del getto non fa aumentare la quantità d'acqua che dal getto può cadere nella tinozza. Ingrandire il rabdoma (o la tinozza) è tuttavia ovviamente il migliore stratagemma possibile per vedere almeno qualcosa quando i fotoni sono talmente rari che solo i grandi fasci (o getti) possono essere percepiti.

Fino a questo punto, nella discussione sull'origine del campo visivo degli insetti basata sul cerchio di confusione è stato necessario evitare i problemi presentati dalla rifrazione (o deviazione della luce) che si verifica quando il fascio luminoso passa dall'aria all'interno dell'occhio dell'insetto. Angoli aventi l'ampiezza del cerchio di confusione Δa (interno all'occhio) e del campo visivo (esterno all'occhio) sono misurati a partire dal punto nodale posteriore, che, in una lente a una faccia come questa, è un punto vicinissimo al centro di curvatura della lente. I raggi luminosi che passano attraverso il punto nodale posteriore all'interno dell'occhio sono paralleli agli stessi raggi all'esterno dell'occhio; per questa ragione il punto nodale posteriore è molto utile per porre in relazione gli angoli misurati all'esterno dell'occhio con gli angoli misurati all'interno. Il diametro di un grande recettore sotteso al punto nodale posteriore è la dimensione $\Delta\varrho$ del campo visivo.

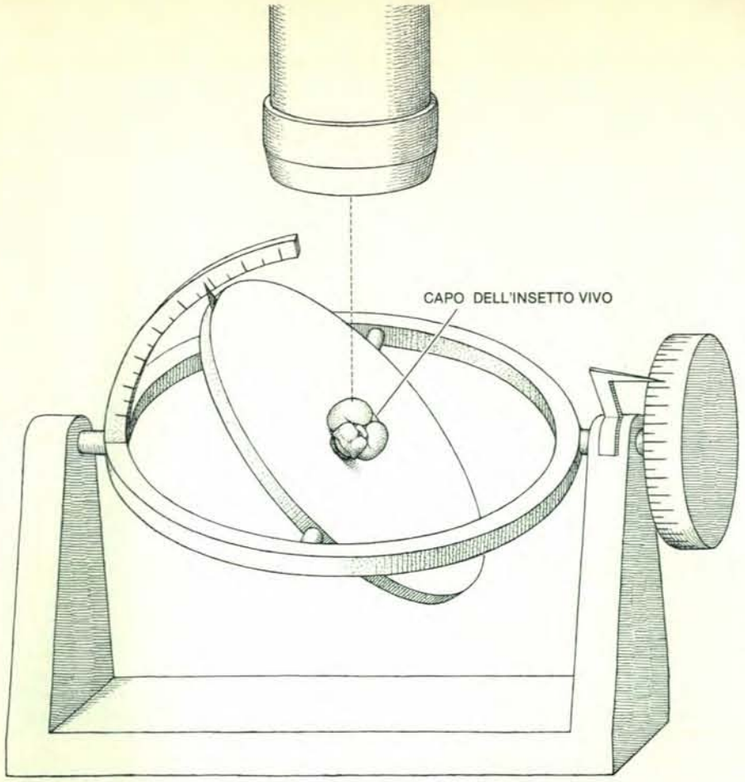
Supponiamo che in un tipico ommatidio ad alto potere risolvente la lente abbia un'apertura di 28,5 micrometri. Se la lunghezza d'onda è di 0,5 micrometri (corrispondente alla luce verde), l'ampiezza Δa del cerchio di confusione è di 1 grado. Se in un ommatidio di questo tipo il diametro del rabdoma (sotteso al punto nodale posteriore) fosse uguale all'ampiezza del cerchio di confusione al 50 per cento d'intensità Δa, allora dalla formula $\Delta\varrho^2 = \Delta a^2 + (d/f)^2$ risulterebbe un campo visivo di 1,4 gradi. Questo rabdoma potrebbe avere un diametro di 1 micrometro e una lunghezza focale f (misurata dal punto nodale posteriore alla punta del rabdoma) di 57 micrometri. Un rabdoma con un diametro inferiore a 1 micrometro non sarebbe molto efficace come rivelatore luminoso e infatti rabdomi così stretti si trovano raramente negli occhi degli insetti. Un'apertura maggiore della faccetta genererebbe un cerchio di confusione più stretto, rendendo necessario, a sua volta, un rabdoma più stretto. Perciò se l'efficienza dell'ommatidio è limitata dalla diffrazione della luce, non ci si dovrebbe aspettare di trovare insetti con ommatidi che hanno faccette più ampie di 30 micrometri circa, perché non si possono avere rabdomi funzionali a una sottigliezza sufficiente per far fronte alle dimensioni delle faccette.



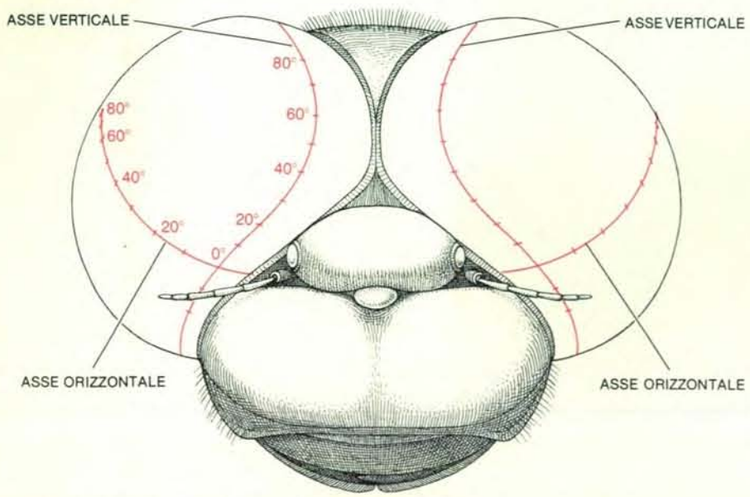
La struttura dell'occhio d'un insetto determina il grado di dettaglio che l'occhio può percepire. Ogni ommatidio ha una cornea che funziona anche come una lente per la luce che viene focalizzata attraverso un cono trasparente su un elemento fotosensibile: il rabdoma. L'asse ottico dell'ommatidio è la linea che si estende attraverso il centro della lente fino al rabdoma. L'angolo interommatidiale $\Delta\phi$ è l'angolo tra gli assi ottici di due ommatidi adiacenti. Le curve in colore sopra ciascun ommatidio indicano di quanto la sensibilità dell'ommatidio decresce con la distanza angolare dell'asse ottico. Il campo visivo dell'ommatidio è definito come l'angolo sotteso dal diametro del campo misurato nel punto in cui la sensibilità è dimezzata rispetto al suo valore massimo.

Tuttavia rimane il fatto che si trovano comunemente faccette di 80 micrometri, particolarmente nella fovea. Alcuni organismi diversi dagli insetti, come il crostaceo stomatopode *Odontodactylus* e lo anfipode di mare profondo *Phronima*, hanno un occhio composto con faccette aventi un diametro fino a 120 micrometri. Tutti gli occhi composti possiedono faccette più ampie e meno numerose di quanto preveda la teoria della diffrazione. Perché si trovano tali grandi aperture? La risposta è che l'aumento del diametro della lente accresce la sensibilità dell'ommatidio proporzionalmente all'area della lente stessa (D^2). La considerazione delle sorgenti puntiformi e del limite di diffrazione non riesce a portarci avanti nella comprensione dell'occhio composto. I singoli ommatidi sembrano essersi adattati in molti modi per catturare anche l'ultimo fotone di luce. Nell'analogia della doccia si deve abbandonare l'idea di cercare di catturare ogni singolo piccolo getto, per cercare invece di distinguere getti a flusso ridotto di tutte le dimensioni e in diverse combinazioni. Quando il flusso di gocce (fotoni) è così ridotto che solo le docce più grandi possono essere rivelate, la sensibilità assume un ruolo importante.

Finora ho concentrato l'attenzione sulle limitazioni ottiche e sulle caratteristiche d'un singolo ommatidio. Tuttavia un occhio composto è formato da centinaia di ommatidi che guardano verso l'esterno in modo più o meno simile in molte direzioni. Il vantaggio di questa disposizione è che l'organo è diviso in settori che operano parallelamente, ossia divide l'informazione proveniente dall'ambiente esterno in molti canali, analizza l'informazione contenuta in ciascun canale e ricombina i dati nei centri nervosi. Così un insetto può essere al corrente in ogni momento dell'intero cam-



La mappa dell'occhio d'un insetto viene preparata montando il capo dell'insetto sulla piattaforma di un goniometro a sospensione cardanica calibrata e fotografando l'occhio sotto angolature diverse. L'angolo necessario per spostare il centro della pseudopupilla di cinque faccette determina l'angolo interommatidiale tra ciascun quinto ommatidio. La posizione dell'asse ottico di ogni quinto ommatidio può in questo modo essere indicata con precisione in una mappa.



Il sistema di coordinate per ricavare la mappa dell'occhio d'un insetto è arbitrario, ma vengono utilizzati i principali assi funzionali dell'occhio. Lo zero dell'asse verticale di solito incrocia lo zero dell'asse orizzontale nella posizione della fovea rivolta in avanti. L'asse verticale zero sulle mappe in coordinate angolari è la sequenza di faccette che guardano nel piano in cui giacciono le direzioni in avanti e verticale; tale asse non è rappresentato sull'occhio da una linea retta. L'asse orizzontale zero sulle mappe è la linea che congiunge le faccette rivolte in direzione orizzontale; non si tratta necessariamente di una fila orizzontale di faccette che attraversa l'occhio, perché gli assi ottici possono essere inclinati rispetto alla superficie dell'occhio.

po visivo che lo circonda, mentre i vertebrati, e quindi anche l'uomo, possono ricevere le informazioni ottiche necessarie solo muovendo in continuazione occhi e capo in una serie di esplorazioni sistematiche successive.

Quali regole determinano la spaziatura degli ommatidi sulla superficie dell'occhio? Nel 1894 A. Mallock (e dopo di lui Horace B. Barlow, Hugo de Vries, Jan Kuiper e molti altri) pensò che la distanza angolare tra gli ommatidi potesse essere calcolata a partire dai limiti posti dalla diffrazione, poiché la distanza angolare di lenti adiacenti ($\Delta\phi$) non deve superare di due volte la distanza ($\Delta\vartheta$) della struttura più fine che può risolvere. Il suo ragionamento era sostanzialmente il seguente.

Il mondo percepito da un insetto può essere considerato come una struttura complessa formata da contrasti tra luce e oscurità. Questa struttura può essere paragonata alla somma di parecchie strutture regolari, costituite da strisce, ciascuna delle quali ha una distribuzione sinusoidale d'intensità tra luce e oscurità e un periodo regolare $\Delta\vartheta$. Affinché un raggruppamento ordinato di recettori veda una struttura a strisce e la ricostruisca con un procedimento parallelo simultaneo, la distanza angolare minima tra i recettori (cioè l'angolo interommatidiale $\Delta\phi$) dev'essere uguale all'angolo sotteso dalla metà della distanza tra una striscia e l'altra (cioè $1/2\ \Delta\vartheta$). Con questa spaziatura minima due picchi (strisce luminose) nell'oscillogramma dell'intensità luminosa, assieme al solco compreso tra loro (striscia scura), si ritrovano su tre recettori adiacenti. Se l'angolo interommatidiale fosse minore, ci sarebbe un numero di recettori più che sufficiente per discernere la struttura; se tale angolo fosse maggiore, ci sarebbero troppo pochi recettori per percepirla. Quest'idea è nota come teorema di Whittaker-Shannon, e in due dimensioni si può applicare a recettori sistemati in una struttura quadrata. Come è stato dimostrato recentemente da Allan Snyder della Australian National University e da William H. Miller della Yale University School of Medicine, una struttura esagonale di recettori, per raggiungere la medesima acutezza visiva di una struttura quadrata, deve possedere più recettori, secondo un fattore pari a 2 diviso per la radice quadrata di 3. Le mappe dell'occhio degli insetti dimostrano che gli assi ottici degli ommatidi si dispongono secondo una struttura che sta fra quella quadrata e quella esagonale. Lo schema strutturale esagonale risulta distorto secondo un modello romboidale particolarmente attorno all'equatore dell'occhio.

Dalla teoria della diffrazione è noto, inoltre, che la distanza $\Delta\vartheta$ fra le strisce nella struttura più fine che possa essere risolta a una qualsiasi lunghezza d'onda della luce attraverso una lente di apertura D è uguale alla lunghezza d'onda λ divisa per l'apertura della lente ($\Delta\vartheta = \lambda/D$). Poiché il minimo angolo interommatidiale $\Delta\vartheta$ dev'essere uguale alla metà

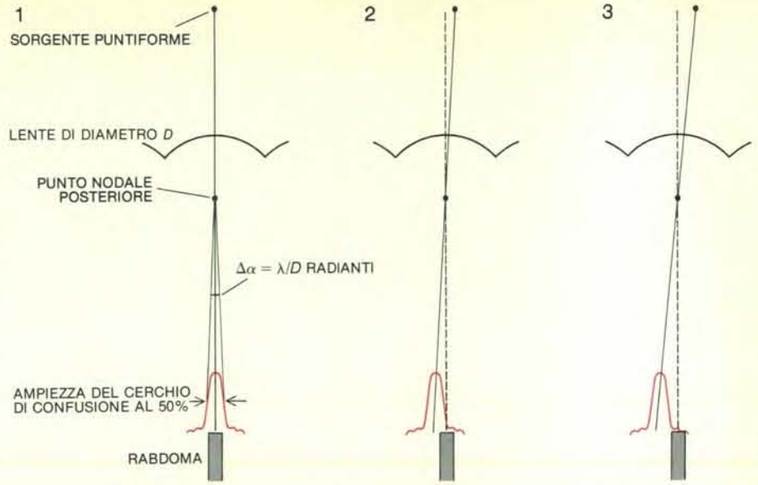
della distanza fra le strisce, affinché un occhio soggetto alle limitazioni dovute alla diffrazione possa vedere, il prodotto dell'apertura della lente per l'angolo interommatidiale dev'essere uguale alla metà della lunghezza d'onda della luce (cioè $D\Delta\phi = \lambda/2$). I primi ricercatori pensavano che le cose stessero proprio così.

Se tuttavia diciamo che l'occhio deve operare a un livello superiore di modulazione luminosa per superare il rumore di fondo dei fotoni, allora possiamo dire che $D\Delta\phi$ è uguale a λ, cosicché $D\Delta\phi$ diventa uguale a 0,5 micrometri. Il prodotto $D\Delta\phi$ dovrebbe perciò essere costante per un occhio che funziona a un dato livello di modulazione, e dovrebbe essere 0,25 micrometri per strutture di assi ad angolo retto, operanti al limite della diffrazione. D e $\Delta\phi$ si possono misurare all'esterno dell'occhio. Sulla base della semplice geometria l'angolo interommatidiale in radianti è il diametro della faccetta D diviso per il raggio ottico R. Perciò D^2/R dovrebbe essere costante per occhi di dimensioni diverse che funzionano alla medesima intensità. In altre parole, conoscendo semplicemente l'apertura D della lente dell'ommatidio e l'angolo interommatidiale $\Delta\phi$, si può determinare una notevole quantità di dati sulla struttura dell'occhio dal prodotto dei due fattori. Ciò si applica sia per insetti diversi sia per regioni diverse dell'occhio d'un singolo insetto.

Se la struttura dell'occhio composto di un insetto fosse fatta in modo da ricostruire i particolari al limite della diffrazione a una lunghezza d'onda di 0,5 micrometri, allora il parametro $D\Delta\phi$ sarebbe fissato a un valore compreso tra 0,25 e 0,31, a seconda di come la struttura degli assi ottici è distribuita tra il modello quadrato e quello esagonale. Tuttavia le mappe oculari rivelano che non vi sono insetti con occhi composti la cui struttura sia determinata da questi limiti. Evidentemente la luce solare non è mai sufficientemente intensa da eliminare il rumore dei fotoni come fattore. Solo le fovee di alcuni insetti attivi durante le ore più luminose del giorno si avvicinano al limite di diffrazione.

Quando una struttura a strisce con una spaziatura di $\Delta\vartheta$ passa attraverso il campo visivo d'un insetto, il rabdoma percepisce la struttura come una modulazione, o oscillazione, applicata sopra una media intensità di luce. Se la spaziatura della struttura striata $\Delta\vartheta$ è grande rispetto all'ampiezza del campo $\Delta\varrho$, l'ampiezza della modulazione sarà piccola. La velocità del movimento non è importante, almeno fintanto che non siano raggiunti livelli ai quali il recettore non riesce a rispondere a tempo.

L'intensità luminosa media dipende dall'illuminazione ambientale delle strisce e dall'ampiezza del campo. L'ampiezza della modulazione M, misurata dal recettore, è in realtà la quantità secondo cui l'intensità luminosa fluttua rispetto all'intensità media. Più precisamente, la modulazione relativa sperimentata dal recettore è uguale alla differenza tra l'intensità massima e quella minima divisa per la somma delle due intensità. Perciò una modulazione spaziale m nel contrasto tra luce e oscurità nell'ambiente visivo viene trasformata dal movimento e dal recettore in una modulazione temporale M, la cui ampiezza relativa dipende dall'ampiezza del campo visivo, dalla distanza fra le strisce e (a velocità significative) dalla velocità angolare del movimento relativo (*si veda l'illustrazione a pagina 114*).



La luce, passando attraverso la lente d'un ommatidio, subisce una diffrazione. Per questa ragione, la luce proveniente da una sorgente puntiforme non viene focalizzata come un punto, ma forma una figura circolare sfuocata: il cerchio di confusione. L'intensità del cerchio di confusione (*curve in colore*) è maggiore al centro e decresce verso la periferia. L'ampiezza del cerchio di confusione ($\Delta\alpha$) è definita come diametro del cerchio dove l'intensità è diminuita del 50 per cento. Quest'ampiezza angolare è uguale alla lunghezza d'onda della luce proveniente dalla sorgente puntiforme divisa per il diametro D della lente. Quando il cerchio di confusione si trova sull'asse ottico (*1*), il rabdoma riceve la maggior quantità di luce. Quando la sorgente puntiforme si sposta dall'asse ottico (*2 e 3*), il cerchio di confusione si sposta dalla punta del rabdoma e la quantità di luce che il rabdoma riceve diminuisce. I raggi sono disegnati in modo da passare per il punto nodale posteriore presso il centro di curvatura della lente; i raggi che attraversano questo punto hanno una traiettoria all'interno dell'occhio parallela a quella seguita all'esterno.

Un recettore con campo visivo $\Delta\varrho$ ampio capterà più luce e possiederà un miglior rapporto segnale/rumore rispetto a un recettore con campo visivo più ristretto (secondo un fattore $\Delta\varrho^2$), ma un recettore con questo campo visivo ampio sarà anche meno sensibile alla modulazione prodotta da strisce strette. Analogamente, un ommatidio con diametro grande della faccetta (D) registrerà una maggiore modulazione assoluta secondo un fattore D^2. È essenziale un trasferimento efficace di modulazione all'interno dell'occhio, particolarmente quando l'intensità della luce ambientale è bassa. Perciò se si potesse determinare il livello del segnale luminoso necessario per attivare i recettori sopra il livello del rumore, si potrebbe arrivare alla relazione ottimale tra il diametro delle faccette e l'angolo interommatidiale.

Una sorgente luminosa estesa di caratteristiche costanti emette un numero medio $\bar{N}$ di fotoni per unità di tempo. Tuttavia i fotoni arrivano al rabdoma in modo irregolare o casuale; in un dato momento, il rabdoma può ricevere un numero di fotoni superiore a quello medio e, nel momento successivo, ne può ricevere un numero inferiore. La variazione del numero effettivo di fotoni rispetto al numero medio è l'inevitabile rumore: questa variazione è definita come deviazione standard rispetto al numero medio di fotoni e, per la luce, è uguale alla radice quadrata di tale numero medio, ossia a $\sqrt{\bar{N}}$.

Il segnale assoluto che il rabdoma riceve effettivamente è l'ampiezza relativa della modulazione M moltiplicata per il numero medio di fotoni captati, cioè $M\bar{N}$. Il rapporto segnale/rumore è perciò $MN/\sqrt{\bar{N}}$, che risulta uguale a $M\sqrt{\bar{N}}$. Questo rapporto segnale/rumore dev'essere abbastanza grande perché il segnale risalti al di sopra del rumore, facendo in modo che l'ommatidio non risponda a falsi allarmi. Perciò, se $\bar{N}$ decresce al limite, M deve aumentare in corrispondenza. Se il rabdoma riceve una media, per esempio, di 4 fotoni per unità di tempo da una sorgente luminosa, il rumore per quel dato numero di fotoni è di $\pm$ 2 fotoni; la modulazione M nel numero medio di fotoni (il segnale) deve essere maggiore del 50 per cento rispetto al numero medio di fotoni, affinché la modulazione risalti al di sopra del rumore. Nella luce intensa, comunque, la modulazione può essere più bassa per tutto il tempo in cui il prodotto $M\sqrt{\bar{N}}$ è ancora di circa 1. Se il rabdoma riceve, per esempio, una

SCHAUM

Una collana di successo per una nuova didattica scientifica

HUGHES-GAYLORD
EQUAZIONI PER L'INGEGNERIA
L. 5.000

McLEAN-NELSON
MECCANICA APPLICATA
L. 8.000

CASHIN-LERNER
RAGIONERIA 1
L. 6.500

SETO
SISTEMI VIBRANTI
L. 6.000

STANSFIELD
GENETICA
L. 7.500

GILES
MECCANICA DEI FLUIDI E IDRAULICA
L. 7.000

TUMA
ANALISI DELLE STRUTTURE
L. 7.500

SPIEGEL
ANALISI DI FOURIER
L. 5.500

SPIEGEL
TRASFORMATE DI LAPLACE
L. 7.000

MASE
MECCANICA DEI CONTINUI
L. 6.500

NASH
RESISTENZA DEI MATERIALI
L. 10.000

SALVATORE
MICROECONOMIA
L. 7.000

AYRES
MATEMATICA GENERALE
L. 10.000

DIULIO
MACROECONOMIA
L. 6.000

VAN DER MERWE
FISICA GENERALE
L. 7.000

HECHT
OTTICA
L. 6.500

CASHIN-LERNER
RAGIONERIA 2
L. 6.500

di prossima pubblicazione

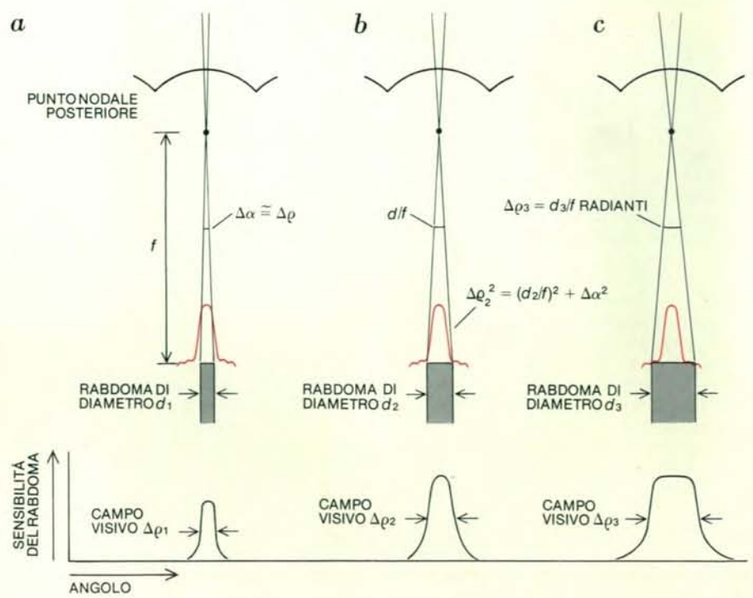
SETO
ACUSTICA
L. 6.000

ETAS LIBRI
Via Mecenate, 87/6 - 20138 Milano

media di 100 fotoni per unità di tempo, il rumore è di ± 10 fotoni, cosicché la modulazione M nel numero medio di fotoni dev'essere solo circa del 10 per cento.

Ci sono ora abbastanza equazioni per calcolare il parametro oculare $D\Delta\phi$ per qualsiasi valore dell'intensità luminosa ambientale: i risultati si possono rappresentare in un diagramma come una curva (*si veda l'illustrazione di pagina 116, ricavata da una recente pubblicazione di Allan Snyder, Doekele Stavenga e Simon Laughlin*). Da questa curva si ricava che il grado di sovrapposizione dei cerchi nella mappa oculare per una visione ottimale può essere calcolato per una qualsiasi intensità luminosa ambientale. Considerando un occhio soggetto ai limiti della diffrazione, il quale deve avere un parametro oculare di 0,3 o meno, i cerchi di diametro λ/D sulle mappe oculari devono sovrapporsi quasi per la metà; l'ampiezza reale del campo $\Delta\varrho$ dovrebbe essere doppia rispetto all'angolo interommatidiale $\Delta\phi$. Per gli occhi di insetti che possiedono un parametro oculare di 0,5 l'ampiezza del campo dovrebbe essere circa 1,3 volte più grande dell'angolo interommatidiale; per quelli con parametro 2 l'ampiezza del campo $\Delta\varrho$ dovrebbe essere solo di 3/4 rispetto all'angolo interommatidiale $\Delta\phi$. Per essere sicuri che queste misure si trovino realmente negli occhi degli insetti, si devono ancora esaminare specie appartenenti a vari gruppi.

La sensibilità dell'occhio d'un insetto può essere aumentata secondo il fattore D^2, aumentando l'apertura D della lente nell'ommatidio o secondo un fattore $(d/f)^2$ diminuendo la distanza f dal punto nodale posteriore al rabdoma: in tutti e due i casi il campo visivo risulta ampliato. Negli occhi composti delle grandi libellule, locuste, mantidi, vespe e coleotteri (e anche negli occhi di molti crostacei) la curvatura della lente è verso la parte interna, cosicché il punto posteriore è un po' più vicino al rabdoma e l'occhio può funzionare ugualmente bene sia che l'animale stia nell'acqua, sia che si trovi fuori dall'acqua. La superficie esterna dell'occhio è poi liscia come il vetro e in molti occhi di insetti c'è un pigmento assorbente che si dispone attorno al rabdoma alla luce intensa. Questo pigmento ha la funzione di assorbire la luce in quegli occhi che debbono funzionare in una vasta gamma d'intensità. Nella luce fioca il pigmento si sposta indietro, aumentando la sensibilità del rabdoma e talvolta allargando il campo visivo dell'ommatidio. In alcuni insetti



La sensibilità e il potere risolutivo del rabdoma sono una funzione del suo diametro e della sua distanza dal punto nodale posteriore. Un rabdoma stretto (*a*) con un diametro d_1, minore dell'ampiezza del cerchio di confusione al 50 per cento d'intensità, riceverà solo una frazione della luce del cerchio, ma reagirà bene alla variazione luminosa attraverso il cerchio di confusione: perciò avrà alto potere risolutivo ma bassa sensibilità. Un ommatidio dotato di rabdoma di questo tipo si definisce «prossimo al limite di diffrazione». Un rabdoma più largo (*b*) con un diametro d_2 pari al doppio dell'ampiezza del cerchio di confusione capterà il 95 per cento della luce del cerchio stesso, ma avrà un campo lievemente più ampio e minore potere risolutivo. Un ommatidio (*c*) con un diametro di rabdoma ancor più grande d_3 ha campo visivo che non è migliore di *b* per le sorgenti puntiformi, ma che è più adatto per svelare oggetti abbastanza estesi in luce fioca. La sensibilità angolare, di ciascun rabdoma è indicata dalle curve in basso. La sensibilità per gli oggetti grandi aumenta con le dimensioni del rabdoma secondo la relazione $(d/f)^2$ così come la pellicola fotografica è più sensibile quando la sua grana è grossa.

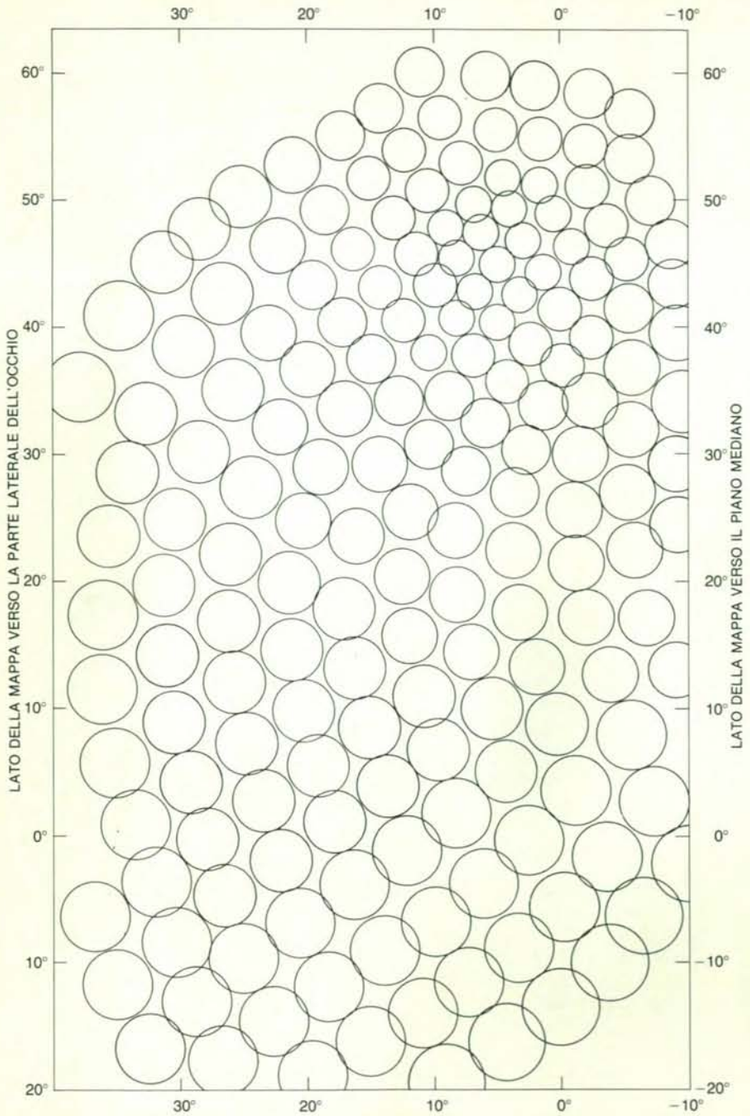
l'adattamento alle basse intensità luminose viene ottenuto mediante il movimento del recettore verso la lente. Gli occhi composti che si adattano in tal modo all'oscurità sono chiamati occhi a cono mobile e si trovano negli emitteri, nelle zanzare e in alcuni coleotteri.

Molti insetti notturni e alcuni lepidotteri e coleotteri diurni hanno occhi con un sistema ottico completamente diverso, col punto nodale posteriore al centro di curvatura dell'intero occhio. La sensibilità di tali occhi è aumentata dal fatto di avere i recettori vicini al punto nodale posteriore. In altri insetti il valore del parametro oculare $D\Delta\phi$ in un punto sull'occhio è completamente diverso dal valore in un altro punto: questo fa pensare che l'occhio sia sensibile a una vasta gamma d'intensità. Vediamo ora in pratica l'importanza di conoscere il parametro oculare per una qualsiasi parte dell'occhio composto, prima di proseguire nello studio dell'occhio in altri modi.

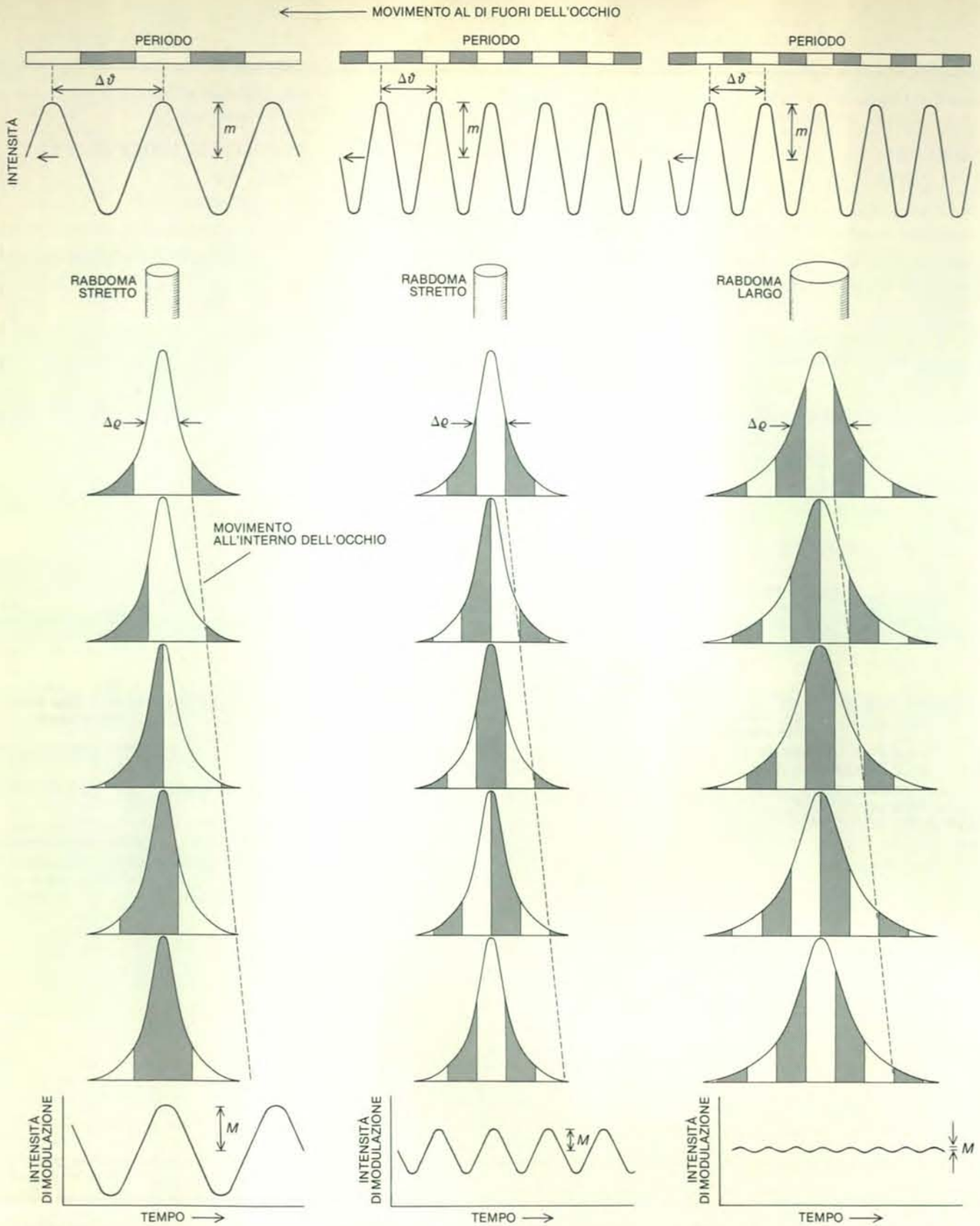
Per esempio, consideriamo l'occhio di una mosca domestica; quando confrontiamo il parametro oculare $D\Delta\phi$ misurato dalle mappe oculari con l'intensità luminosa ambientale in cui la mosca mostra di solito la sua attività, notiamo immediatamente quella che sembra essere una discrepanza: l'occhio della mosca è apparentemente destinato a operare con un'intensità luminosa circa 10 000 volte più bassa rispetto a quella della piena luce solare. La discrepanza si risolve quando ci si ricorda che le mosche non solo preferiscono i luoghi ombrosi, ma usano gli occhi anche compiendo giravolte ad alta velocità. A causa di queste manovre, l'ambiente visivo passa davanti all'occhio a velocità angolare molto elevata, riducendo il numero reale di fotoni che cadono su un singolo fotorecettore: le giravolte compiute ad alta velocità diminuiscono l'intensità luminosa percepita di circa 1000 volte. Queste considerazioni spiegano anche perché i sirfidi e le libellule si librano, mentre le mantidi rimangono immobili.

Tenendo presente l'ottica dell'occhio composto ci proponiamo ora di occuparci di alcune mappe di occhi d'insetti preparate mediante misurazioni della pseudopupilla. Ogni cerchio delle mappe oculari ha un diametro di $5\lambda/D$ radianti, con la lunghezza d'onda λ della luce verde, e cioè pari a 0,5 micrometri. Ogni cerchio rappresenta l'ampiezza del cerchio di confusione di ogni quinto ommatidio. La dimensione dei cerchi è perciò la chiave per la comprensione del potere risolutivo di ciascun ommatidio: quanto più piccolo è il cerchio, tanto migliore è la risoluzione. La distanza dal centro d'un cerchio al centro del successivo corrisponde a cinque volte l'angolo interommatidiale ($5\Delta\phi$). La vicinanza dei cerchi fornisce la misura della densità con cui l'occhio dell'insetto raccoglie i dati del mondo visivo e perciò rappresenta la risoluzione visiva in quella regione dell'occhio. Il sovrapporsi dei cerchi indica il valore del parametro oculare $D\Delta\phi$ in quel punto» Per esempio, dove i cerchi si toccano appena, il loro diametro λ/D è uguale alla loro separazione $\Delta\phi$ e il parametro oculare ha un valore di 0,5.

Ho compilato mappe degli occhi di alcune specie rappresentative d'insetti e ho rilevato un certo numero di caratteristiche generali. Per prima cosa, non sembrano esistere insetti i cui occhi siano adattati principalmente a funzionare vicino al limite di diffrazione. Sono errate le valutazioni precedenti, secondo cui gli occhi degli insetti sarebbero limitati dalla diffrazione. Persino gli occhi della vespa



Questa mappa oculare mostra la parte frontale dell'occhio destro della libellula *Austrogomphus*, che possiede una fovea che guarda obliquamente in avanti e verso l'alto. Ogni cerchio rappresenta il campo più stretto possibile di ciascun quinto ommatidio. Per compensare il fattore cinque, il diametro è uguale a cinque volte la lunghezza d'onda della luce verde (pari a 0,5 micrometri) diviso per il diametro della lente dell'ommatidio, e cioè $5\lambda/D$. I cerchi più piccoli indicano che le faccette in quell'area dell'occhio sono più grandi e hanno un potere risolutivo maggiore rispetto a quelle situate in zone con cerchi più ampi. Il centro di ogni cerchio rappresenta la posizione dell'asse ottico di ciascun quinto ommatidio; la distanza dal centro di un cerchio al centro del successivo è perciò uguale a cinque volte l'angolo interommatidiale, ossia $5\Delta\phi$. La misura della distanza o della sovrapposizione dei cerchi fornisce il valore del parametro oculare $D\Delta\phi$ che è una misura critica di quanto dista dal limite di diffrazione l'occhio dell'insetto in ciascuna zona. In questo occhio solo un quarto delle faccette si troverebbe al limite della diffrazione. Il centro della fovea è al centro del raggruppamento di cerchi piccoli che si trovano prossimi alle coordinate 0 e 50 degli assi situati nello spazio angolare.



Una struttura a strisce scure e chiare (*in alto*) che si sposta attraverso il campo visivo d'un ommatidio viene percepita dal rabdoma come una modulazione dell'intensità luminosa nel tempo. Le strisce hanno una distribuzione sinusoidale dell'ampiezza *m* dalla luce al buio (*curve nella seconda fila dall'alto*). La quantità di modulazione percepita dal rabdoma (*strutture cilindriche nella terza fila dall'alto*) è funzione del campo visivo del rabdoma $\Delta\varrho$ e della distanza tra le strisce $\Delta\vartheta$. A sinistra, un rabdoma stretto guarda verso un insieme di larghe strisce in movimento; il campo di visuale del rabdoma è un quarto del periodo delle strisce. La sensibilità del rabdoma attraverso il suo campo visivo è rappresentata dalla forma delle curve disegnate sotto la struttura cilindrica del rabdoma. Quando le strisce larghe attraversano lo stretto campo visivo del rabdoma, questo percepisce una modulazione di grande ampiezza *M* (*curva in basso a sinistra*). Nella colonna al centro, un rabdoma stretto guarda verso una struttura di strisce strette in movimento e percepisce una modulazione più rapida, di ampiezza modesta. Nella colonna a destra, un rabdoma ampio guarda verso una struttura di strisce strette in movimento e vede una figura opaca dotata di rapido sfarfallamento, che potrebbe essere al di sotto del livello del segnale di disturbo ottico a meno che l'illuminazione sia molto intensa o le faccette dell'occhio molto grandi. Le linee tratteggiate diagonali indicano il movimento all'interno dell'occhio.

della sabbia australiana *Bembix*, che conduce vita attiva sotto la luce accecante del sole nel deserto, hanno un parametro oculare di 0,32 solo nel centro della fovea. Gli occhi delle mantidi, che in genere catturano le prede nella luce intensa, hanno un parametro oculare sorprendentemente alto, variabile fra 0,5 e 2.

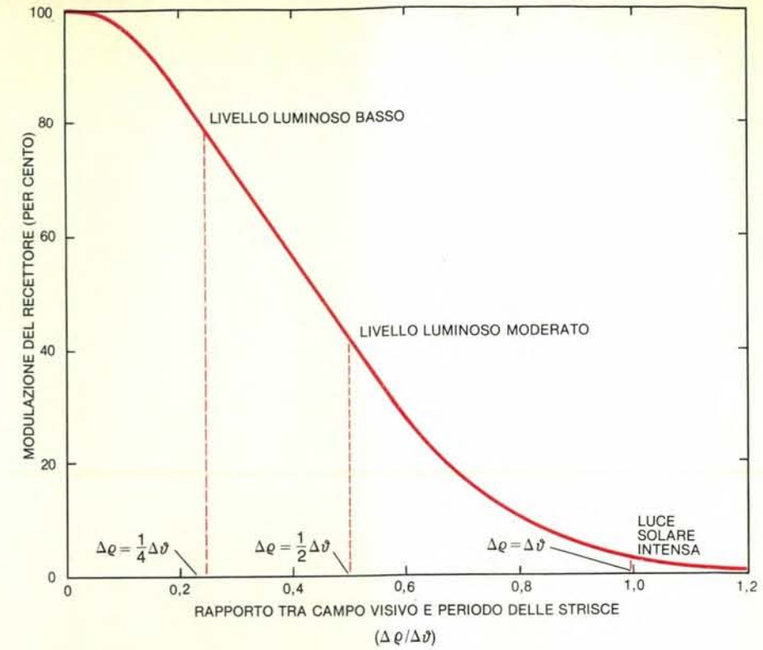
In secondo luogo, la maggior parte degli insetti che volano in piena luce solare, come api, farfalle e libellule, possiede un parametro oculare di circa 0,5. Questa generalizzazione dovrebbe essere usata con cautela, perché una piccola variazione del valore del parametro oculare intorno a 0,5 implica una differenza notevole nell'intensità luminosa ottimale alla quale l'occhio può funzionare nelle condizioni migliori.

In terzo luogo, gli occhi degli insetti carnivori sembrano avere un parametro oculare maggiore rispetto agli occhi degli insetti erbivori presenti nel medesimo habitat: anche l'insetto carnivoro possiede tuttavia una fovea.

In quarto luogo, è abbastanza frequente trovare un lieve gradiente nel valore del parametro oculare che attraversa orizzontalmente l'occhio d'un singolo insetto a partire dai valori minimi presso la parte frontale a quelli massimi verso la parte laterale. Presumibilmente i gradienti nel parametro oculare sono in rapporto col fatto che, quando l'insetto vola, gli oggetti ai lati dell'occhio hanno una velocità angolare maggiore e producono meno modulazione rispetto agli oggetti che stanno di fronte all'occhio. I gradienti sono più facilmente osservabili seguendo una fila orizzontale di faccette negli occhi delle locuste, degli imenotteri e delle libellule, ma sono caratteristici anche delle mantidi, delle formiche e di altri insetti che non hanno file orizzontali di faccette o non presentano movimento rapido in avanti.

In quinto luogo, lungo la regione equatoriale dell'occhio la pseudopupilla è allungata verticalmente, perché qui l'angolo interommatidiale tra le faccette nelle file verticali è minore rispetto all'angolo interommatidiale tra le faccette nelle file orizzontali. Benché la struttura esagonale delle faccette sull'occhio risulti poco modificata quando la pseudopupilla è allungata, gli assi ottici delle faccette sono compressi nella direzione verticale, per formare un reticolo tipo diamante, cosicché gli occhi scrutano l'ambiente esterno da una serie ordinata di faccette romboidali. Come abbiamo visto, questa struttura a quattro lati permette all'occhio di raccogliere dati dell'ambiente esterno in modo più completo rispetto a quanto accadrebbe in una struttura esagonale. Queste considerazioni suggeriscono la possibilità che gli occhi usino le file verticali di ommatidi piuttosto che quelle orizzontali per compiti che richiedono alta risoluzione.

Infine, dove esiste notevole sovrapposizione binoculare, l'occhio esplora un campo visivo più ampio e il parametro oculare è quindi sempre maggiore di quanto si pensi. Questo è il caso delle mantidi e degli zigotteri (un gruppo di odonati).



L'ampiezza della modulazione relativa *M* percepita dal rabdoma quando una struttura a strisce si sposta attraverso il suo campo visivo è rappresentata in funzione del rapporto tra l'ampiezza del campo e la distanza tra le strisce. Quando il campo è uguale alla distanza delle strisce ($\Delta\varrho = \Delta\vartheta$), la modulazione è pari a poche unità per cento, ma può essere sufficiente per vedere nella piena luce solare. Quando il campo è uguale a metà della distanza ($\Delta\varrho = 1/2\ \Delta\vartheta$), la modulazione è di circa il 40 per cento e sarebbe sufficiente a livelli luminosi bassi per superare il disturbo causato dai fotoni. Quindi i recettori adatti a operare a livelli luminosi bassi possiedono un ampio campo visivo ($\Delta\varrho$) e, di conseguenza, debbono avere un grande diametro (*d*).

Dalle mappe oculari si possono ricavare anche molte informazioni sulle fovee degli insetti. Si può, per esempio, notare che una fovea si forma dove file di assi ottici sono raggruppate ed è quindi maggiore la densità degli ommatidi. Una fovea non si forma mai per interpolazione di file soprannumerarie di ommatidi. È vero anzi il contrario: nella regione foveale alcune file di faccette debbono essere eliminate per lasciare il posto a faccette di maggiori dimensioni.

Sulle mappe oculari di alcuni insetti predatori si può osservare che i cerchi di diametro λ/D si sovrappongono sui bordi più nella fovea che nelle regioni adiacenti dell'occhio. Questo significa che nella fovea il parametro oculare decresce, il che a sua volta indica che in generale la fovea è una regione che si è adattata alla visione in luce relativamente intensa. Tuttavia in taluni insetti, come la libellula *Austrogomphus*, il parametro oculare rimane costante attraverso la fovea.

Le file di faccette sono sempre più regolari nella fovea di quanto siano in altre zone dell'occhio: questo dimostra che la regolarità ha un ruolo importante in quelle regioni dell'occhio dove è necessaria la vista migliore possibile.

Benché le faccette nella fovea siano di solito più grandi di quelle delle altre zone nell'occhio, questo fatto non si verifica negli occhi di *Locusta* e della mantide *Orthodera*. Entrambi questi insetti hanno una fovea che guarda in avanti, in cui l'angolo interommatidiale è minore e in cui le faccette hanno le stesse dimensioni di quelle che si trovano in altre zone dell'occhio. Questa fovea è forse un'area specializzata dove l'acutezza visiva risulta aumentata, ma dove i campi sono più ristretti e perciò meno sensibili.

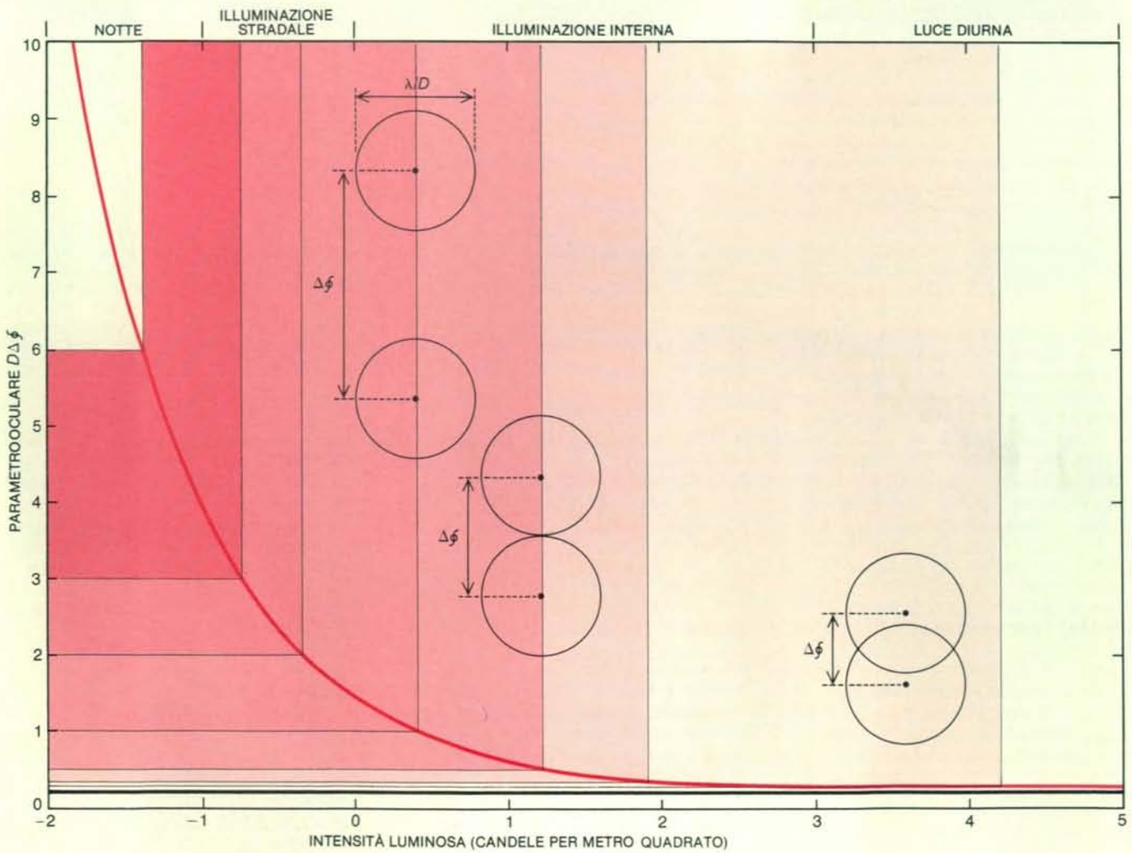
Nella vespa della sabbia *Bembix*, che si sposta nella luce accecante del deserto australiano, il parametro oculare è basso, circa 0,31 nella fovea, vicino al limite di diffrazione. Molti insetti tuttavia possiedono un parametro oculare notevolmente alto persino nella fovea. Per esempio, il parametro oculare raggiunge il valore di 1 nella fovea degli zigotteri (un gruppo di odonati), che ricercano la preda tra le canne di palude e sotto gli alberi. Benché in ogni caso la fovea sia una regione avente potere risolutivo più elevato, è chiaro che può essere una regione di bassa sensibilità o può, avendo faccette più grandi, essere destinata a operare alla medesima intensità luminosa del resto dell'occhio.

Le abitudini visive degli insetti, possiamo esserne certi, stanno scritte sui loro globi oculari; purtroppo è assai dif-

ficile scoprire indicazioni precise perché sappiamo molto poco sul modo con cui essi usano gli occhi. Presumibilmente gli occhi che guardano lateralmente, come quelli di alcuni uccelli o del coniglio, permettono di segnalare il pericolo con anticipo. Un indizio è fornito dalle locuste e dalle mosche, che sono molto sensibili ai movimenti che si compiono ai lati del capo e sono relativamente insensibili ai movimenti che avvengono davanti a loro. Le grandi fovee di molte grosse libellule, puntate direttamente in senso laterale, sono tuttora un enigma insoluto: forse servono all'animale per librarsi sopra le acque; gli angoli interommatidiali tra le faccette della fovea misurano meno di 0,5 gradi e sono i più piccoli che mi sia mai accaduto di trovare tra gli insetti.

Gli insetti che cacciano mediante la vista presumibilmente hanno bisogno di una fovea che guardi in avanti. Tuttavia non si sa ancora con certezza se la maggior parte delle fovee degli insetti, come quelle dei vertebrati, funzionino meglio in piena luce: se ciò fosse vero, le libellule e le mantidi sarebbero soggette a una diminuzione dell'acutezza visiva al crepuscolo, più o meno come avviene negli esseri umani.

Gli occhi delle mantidi presentano ancora parecchi problemi da risolvere. In molte mantidi la fovea è evidentissima. Eppure nel 1971 J.C. Barros-Pita e H. Maldonado in Venezuela hanno dimostrato che l'area immediatamente circostante la fovea, come viene definita in questo articolo (non quindi il centro della fovea), è essenziale per permettere all'insetto di vedere abbastanza bene per poter catturare le sue prede. Inoltre in tutta una serie di mantidi da me scelte, di cui ho tracciato le mappe oculari, nessuna aveva cerchi di diametro λ/D che si sovrapponessero al di fuori della regione della fovea. Fatto ancor più notevole, gli occhi delle mantidi di maggiori dimensioni, che possedevano faccette aventi diametro fino a 50 micrometri al centro della fovea, avevano un parametro oculare in altre parti maggiore di 2; perciò la separazione da un centro all'altro dei cerchi sulle mappe oculari è più che doppia rispetto al loro diametro λ/D. L'unica spiegazione possibile è che la mantide abbia 45 gradi di sovrapposizione binoculare su ciascun occhio: gli occhi risultano puntare quindi sopra un campo visivo eccezionalmente ampio.



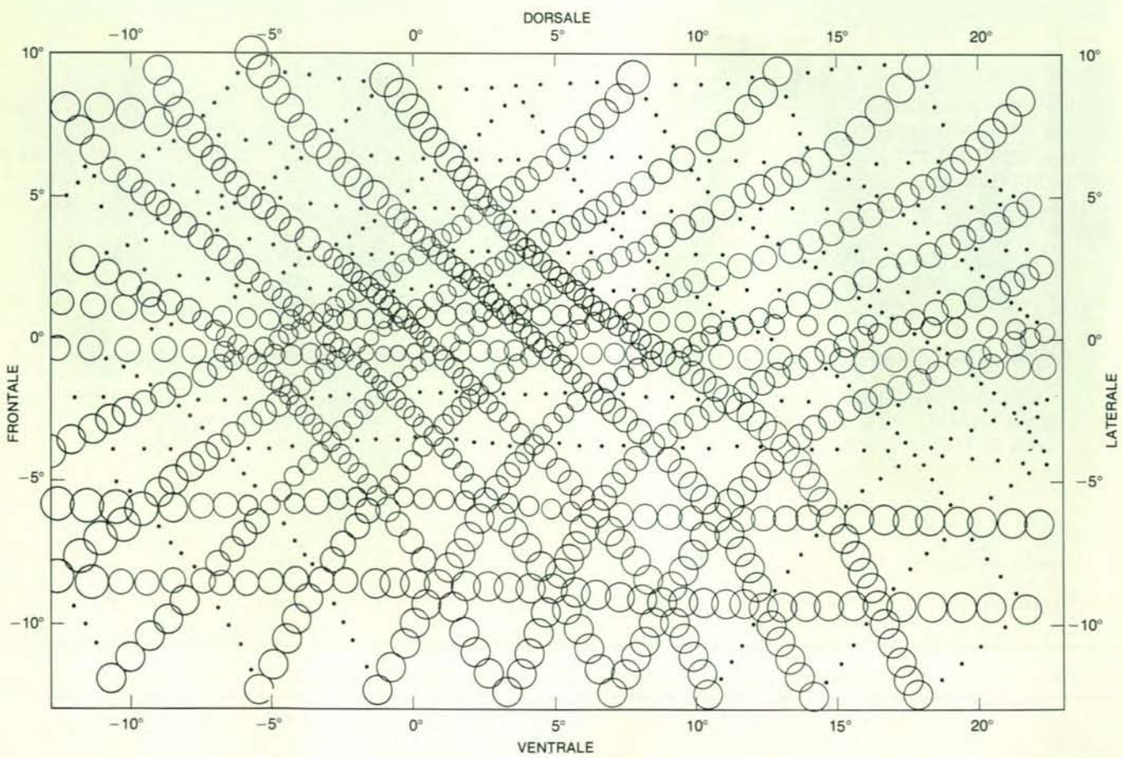
Il parametro oculare $D\Delta\phi$ (*curva in colore*) ha un valore ottimale diverso per differenti livelli di intensità luminosa ambientale. Se la struttura degli occhi degli insetti fosse determinata solo dalla limitazione legata alla diffrazione della luce, gli occhi di tutti gli insetti avrebbero un parametro oculare di circa 0,25 (*linea orizzontale in neretto, attraverso la parte bassa del diagramma*). I cerchi di λ/D sulle mappe oculari in tal caso si sovrapporrebbero per metà (non indicato in figura). Ma in realtà solo nelle fovee degli insetti attivi con luce solare estremamente intensa il parametro si avvicina al limite di diffrazione (*a destra*). Inoltre, l'occhio d'un insetto spesso presenta valori differenti di parametro oculare nelle diverse zone, ovviamente perché ciascuna area è più efficiente al livello d'intensità luminosa in cui viene più usata. L'unità mediante la quale è espressa l'intensità luminosa (*asse orizzontale*) è la candela per metro quadrato.

A prima vista, anche il fatto che le libellule abbiano fovee che guardano obliquamente in avanti e verso l'alto è sconcertante. Tuttavia si può osservare che un elicottero che vola a tutta velocità si inclina in avanti, in modo che la portanza delle sue pale rotanti fornisca anche la spinta in avanti. Le libellule hanno una certa rassomiglianza con gli elicotteri, ed è possibile che le loro fovee, rivolte verso l'alto, guardino in realtà diritto verso la preda che stanno inseguendo in volo. Inoltre, una fovea che guarda verso l'alto è esattamente ciò di cui un insetto ha bisogno allo scopo di anticipare la posizione di un oggetto in caduta o di avvicinarsi a una preda dal basso. In più, la parte superiore dell'occhio di libellula è particolarmente sensibile alle lunghezze d'onda brevi, il che starebbe a indicare che l'occhio percepisce il migliore contrasto e possiede la migliore risoluzione quando guarda gli oggetti sullo sfondo del cielo azzurro.

La relazione teorica tra il valore del parametro oculare e l'effettiva intensità di luce in cui ciascuna parte dell'occhio composto deve funzionare serve per raccogliere una notevole messe di dati. Alla base della teoria stanno tre ipotesi fondamentali: primo, che l'occhio composto è un sistema ordinato di recettori che raccolgono simultaneamente l'informazione attraverso canali paralleli; secondo, che le faccette e i rabdomi dell'occhio devono essere abbastanza grandi per captare un numero sufficiente di fotoni e percepire così la modulazione luminosa sopra il rumore; terzo, che una struttura a strisce risolta da un qualsiasi singolo recettore può essere anche percepita come struttura da tutto l'occhio nel suo complesso.

In base a una rassegna di studi sulla visione degli insetti si potrebbe sollevare l'obiezione che esistono poche prove sulla verità dell'ultima ipotesi, ossia che una struttura fine possa essere ricostruita effettivamente da un qualsiasi occhio composto. Tuttavia le prove esistono. Per prima cosa, dopo numerosi test compiuti sulle api, esercitate a rispondere a modelli visivi, si è potuto dimostrare che due sono gli aspetti del modello importanti per un'ape: lo sfarfallamento che il modello crea quando passa veloce attraverso la visuale dell'occhio composto e gli angoli e i colori dei bordi contrastanti del modello; modelli che sembrano assai differenti al nostro occhio vengono facilmente confusi da un'ape esercitata, se tali modelli producono lo stesso sfarfallamento. In secondo luogo, affinché un insetto percepisca la velocità e la direzione di un modello che si sposta rispetto al suo occhio, l'aspetto più importante del segnale non è il modello in sè, ma l'impressione sequenziale del segnale che attraversa l'occhio.

È difficile escogitare un esperimento che dimostri in modo decisivo se un insetto ricostruisce o no un'immagine nei suoi centri nervosi, percependola come una vera e propria struttura. Invece è facile dimostrare che un insetto può reagire a un piccolissimo movimento di un grosso oggetto e che risponde particolarmente bene anche a un minimo movimento d'un piccolo oggetto. In molti insetti di piccole dimensioni, che non hanno nel capo spazio per grandi occhi, l'angolo interommatidiale può essere ampio fino a 10 gradi: eppure con un certo numero di esperimenti accurati si è potuto dimostrare che questi insetti possono reagire a un oggetto particolare che sottende un angolo pari a un decimo dell'angolo interommatidiale. Parte del segreto sta nel fatto che alcuni insetti



La mappa oculare della vespa della sabbia, *Bembix palmata*, rivela che gli assi ottici degli ommatidi dell'insetto (*puntini e centri dei cerchi*) non seguono la medesima struttura esagonale delle faccette dell'occhio. Lungo l'asse orizzontale della mappa i punti sono disposti in un reticolo romboidale che diventa sempre più compatto man mano che si avvicina all'equatore dell'occhio. I cerchi hanno un diametro di λ/D. Le sovrapposizioni marcate tra un cerchio e l'altro indicano che il parametro oculare si avvicina a 0,3. *Bembix* è uno dei pochi insetti per cui il parametro oculare (e solo nella fovea) si avvicini di parecchio al limite di diffrazione. *Bembix* vive nella luce accecante del deserto australiano, cosicché il segnale di disturbo dei fotoni è basso rispetto all'intensità della luce e il limite di diffrazione può essere avvicinato.

ZANICHELLI NOVITÀ

BATTISTINI
LETTERATURA E SCIENZA
L. 3.200

HINDE
BASI BIOLOGICHE DEL COMPORTAMENTO SOCIALE UMANO
L. 9.800

DEMENT
E C'È CHI VEGLIA E C'È CHI PUÒ DORMIRE
Introduzione allo studio del sonno e dell'attività onirica.
L. 5.200

OLIVERIO FERRARIS
IL BAMBINO IN CASA E A SCUOLA
L. 3.200

DE ROBERTIS, SAEZ, DE ROBERTIS JR.
BIOLOGIA DELLA CELLULA
3.a edizione italiana.
L. 14.200

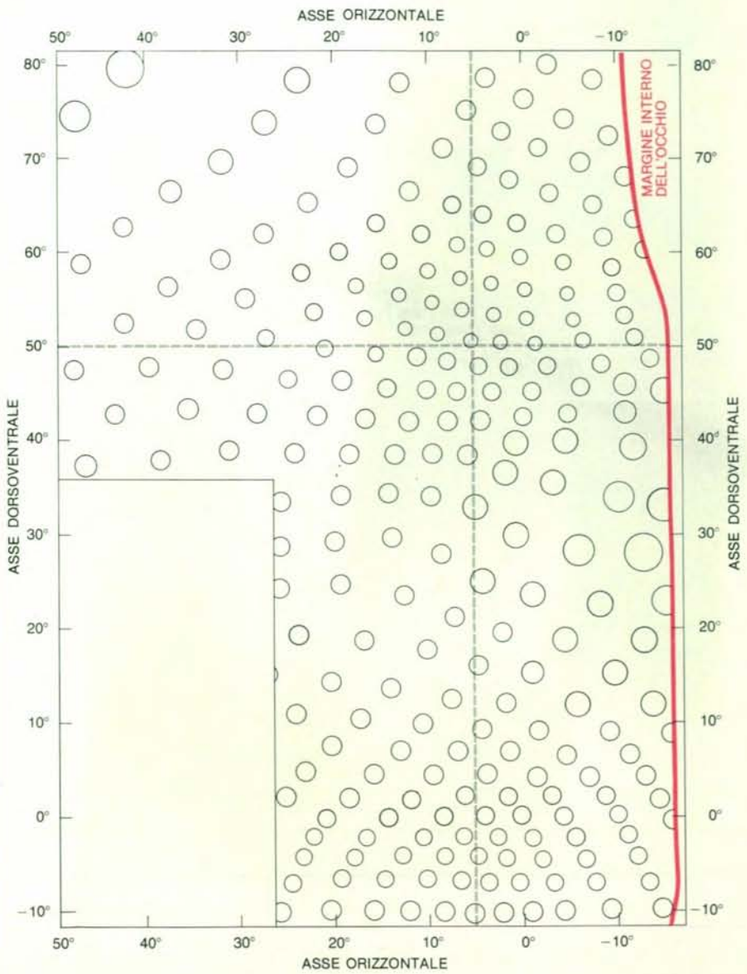
GREEN
INTRODUZIONE ALLA FISIOLOGIA UMANA
L. 7.800

GARDNER
CARNEVALE MATEMATICO
Da Scientific American nuovi problemi divertenti di logica e varia scienza
L. 6.400

KLEPPNER, RAMSEY
ELEMENTI DI ANALISI MATEMATICA
secondo il metodo dell'istruzione programmata.
L. 4.000

possono far uso del tremore oculare per esaminare l'ambiente visivo e ricostruire un quadro più particolareggiato in una struttura temporale anziché spaziale. Tale affermazione andrebbe contro l'ipotesi che la raccolta di dati dev'essere simultanea.

Fino a un certo punto, le obiezioni che si possono sollevare contro le ipotesi discusse in questo articolo si possono annullare a vicenda. Per esempio, il fatto di vedere il movimento richiede che l'insetto sia sensibile all'esatta sincronizzazione di piccole modulazioni in ciascun asse visivo sopra una regione dell'occhio. Il fatto di vedere la direzione di un margine contrastante richiede che l'insetto sia sensibile all'esatta fase dei componenti sinusoidali del modello visivo. Il fatto di vedere il movimento relativo di un piccolo oggetto scuro in una direzione imprevedibile su uno sfondo a chiazze dal disegno complicato, che si muove a sua volta, è una realtà che si presenta spesso a un insetto predatore che vola, per cui il suo occhio deve raccogliere tutte le informazioni possibili. Penso che quando esamineremo più attentamente il comportamento visivo degli insetti, troveremo che essi ricavano dei benefici dalla raccolta di ogni possibile informazione. La forma degli occhi composti più progrediti suggerisce l'ipotesi che essi si siano adattati in modo da poter registrare la massima quantità di informazioni possibile, in una specie di continua battaglia contro il disturbo provocato dai fotoni.



Questa mappa oculare della farfalla notturna *Zyxomma* mostra l'asse ottico di un ommatidio ogni cinque. *Zyxomma*, un insetto dell'Asia sudorientale, si sposta in cerca di cibo solo al crepuscolo. La mappa indica l'estrema separazione tra i cerchi (aventi diametro $5\lambda/D$ in radianti), la fovea piana dell'occhio dell'insetto, che guarda in avanti e verso l'alto presso le coordinate angolari 0° e 50°, e la densità elevata di assi ottici lungo l'asse orizzontale zero (*in basso nella mappa*). Con il tratteggio sono indicate le linee lungo le quali non vi sono errori di distorsione provocati dallo sviluppo della superficie curva dell'occhio dell'insetto su di un piano.

GIOCHI MATEMATICI

di Martin Gardner

Dividere in parti uguali è un procedimento che conduce a interessanti settori della matematica

«Il Mago non perse tempo, ma, balzando in avanti, sollevò la spada affilata, la fece roteare una o due volte sopra la testa e quindi vibrò un colpo possente che tagliò il corpo dello stregone esattamente in due parti.

Dorothy urlò...»

- L. Frank Baum
Dorothy and the Wizard in Oz

Un gioco che si trova spesso nei vecchi libri di rompicapo è quello di dividere una forma data in due, tre o più parti uguali. Uguali significa talora congruenti, talora semplicemente uguali di area. Coloro che leggono questa rubrica da parecchio tempo ricorderanno molti problemi di questo tipo: calcolare in quanti modi si può dividere una scacchiera in due o quattro parti congruenti seguendo le linee del reticolato, dividere il simbolo dello yin-yang (con una linea retta) in quattro parti di area uguale, dividere una torta quadrata in n spicchi di uguale volume, *rep-tile* che si possono dividere in copie congruenti di se stessi e numerosi altri giochi.

Questo mese esamineremo problemi di divisione in parti uguali che non sono stati ancora presi in considerazione. Alcuni di essi conducono a settori importanti della matematica moderna.

Iniziamo con il problema più semplice: dividere una forma piana in due parti congruenti. I pezzi si considerano congruenti se sono immagini speculari uno dell'altro. Si potrebbe pensare che questi problemi siano facili da risolvere; in realtà possono essere noiosi e difficili. Per quanto ne so, non esiste un algoritmo per decidere, in generale, se una forma si può dividere in due o più parti congruenti e, stranamente, sono scarsi i teoremi interessanti su tali divisioni.

Il lettore è invitato a mettere alla prova la sua abilità dividendo ciascuna delle 12 forme mostrate nella figura della pagina a fronte in due parti identiche. Non sono permessi trucchi: ogni metà deve essere semplicemente connessa (non ci devono essere buchi o parti unite da singoli punti). Alcune di queste figure sono riprese dalla rubrica di giochi matematici di una rivista francese tenuta da Pierre Barloquin. (Scribner's ha di recente pubblicato le traduzioni di tre popolari libri di rompicapo di Barloquin.) La forma con il buco è difficile e introduce in una regione inesplorata dei rompicapo: quella delle forme con buchi che devono essere divise in parti congruenti.

Si noti che la quarta forma possiede solo una simmetria bilaterale, dato che l'asse di simmetria passa verticalmente a metà. Ogni figura piana o solida con simmetria bilaterale si può ovviamente dividere in due parti congruenti tagliandola lungo il suo asse o piano di simmetria. (La figura a pagina 124 mostra come il mago di Oz abbia tagliato in due parti congruenti un crudele Mangaboo sotto gli occhi di Eureka, il micio di Dorothy. La scena si svolge nella Città di Vetro, sotto la superficie terrestre, i cui abitanti sono vegetali.) È sempre necessario tagliare lungo un asse o un piano di simmetria per dividere una figura a simmetria bilaterale in parti congruenti? La risposta è no, come dimostra la quarta forma. Si chiede quindi di trovare per questa figura una soluzione che sia diversa da quella più immediata.

Dividere una forma in n parti congruenti diventa generalmente più difficile col crescere di n, specialmente se viene specificata la forma delle parti. I 12 pentomini dell'illustrazione in alto a pagina 125 pongono un interessante problema di quadrisezione. (Devo ricordare che il nome «pentomini» fu registrato come marchio di fabbrica nel 1975 da Solomon W. Golomb, che coniò il termine.) Quanti pentomini si possono dividere in quattro parti congruenti? Tutti meno tre. Scoprii con sorpresa che i nove con soluzione potevano tutti essere risolti usando la stessa forma componente, un pentomino più piccolo, e che questa forma era unica. È capace il lettore di scoprire e identificare i tre pentomini impossibili?

Lasciamo cadere il requisito della congruenza, limitandoci a richiedere che le n parti abbiano area uguale. Appare intuitivamente chiaro che ogni forma piana si può tagliare con una linea retta in due metà di area uguale. Non è forse così immediatamente ovvio che la retta può essere parallela a qualsiasi retta data esterna alla figura. Per dimostrare questa relazione ci serviamo di un famoso teorema: una funzione continua in una variabile, definita su un intervallo chiuso e limitato di estremi A e B, ha un valore minimo e uno massimo e assume tutti i valori reali compresi tra questi due.

Facciamo un esempio: state camminando in montagna lungo un sentiero tortuoso che va da A a B. L'altitudine a cui vi trovate a ogni momento è una funzione continua della vostra posizione sul sentiero. Il teorema ci dice che sul sentiero c'è almeno un punto (ce ne può essere più di uno se il sentiero sale e scende) di altitudine minima, almeno un punto di altitudine massima e almeno un punto per ogni valore reale dell'altitudine compreso tra questi due. Il teorema sembra così ovvio da essere banale eppure si rivela di una straordinaria potenza nel dimostrare teoremi che non sono affatto ovvi.

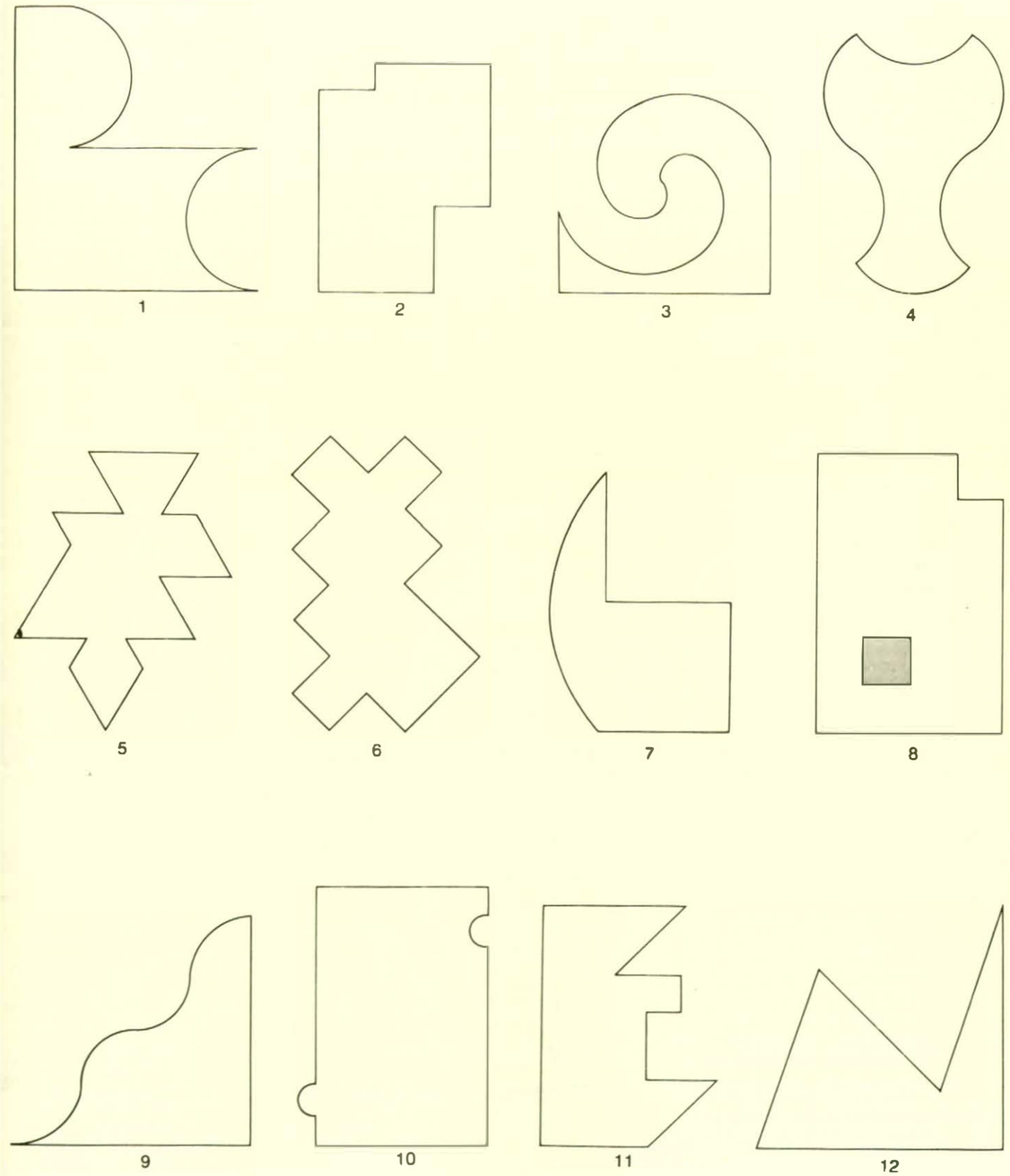
Si consideri la regione scura a sinistra nella figura in mezzo di pagina 125. All'esterno c'è una retta arbitraria x. Vogliamo dimostrare che, all'interno della figura, si può tracciare una retta parallela a x che divida l'area esattamente in due parti uguali. Si immagini una retta che si muova lentamente, sempre parallelamente a x, nella direzione perpendicolare indicata dalla freccia tratteggiata. La retta tocca la regione in A e la abbandona in B. Dato che la retta si muove sulla regione, l'area al di sotto di essa è una funzione continua della distanza che essa ha percorso a partire da A. Tale area è zero in A e massima in B. Per il nostro teorema tra A e B esiste un punto in cui l'area è esattamente la metà del massimo, cioè un punto in cui la retta divide esattamente in due parti la area.

La dimostrazione è così generale che si applica non solo a qualsiasi forma connessa, comprese quelle con buchi, ma anche a regioni sconnesse. Basta uno sguardo alla parte destra della figura in mezzo di pagina 125 per convincersi che si può tracciare una retta parallela a una retta data attraverso un numero qualsiasi di regioni in modo tale che l'area totale da una parte della retta sia uguale all'area totale dell'altra.

Si supponga di avere due regioni di tipo qualsiasi nel piano, come si vede nella figura in basso a pagina 125. È sempre possibile tracciare una retta che divida contemporaneamente ogni regione in parti di area uguale? Possiamo dimostrare che è possibile. Si divida prima in due parti uguali la regione bianca con una retta che non incontra quella grigia. Si faccia ruotare la retta, sempre conservando la bisezione della regione bianca. (Sappiamo che questo è possibile per il teorema precedente.) Come si vede dalla figura, la retta ruotante toccherà la regione grigia nel punto A e ne uscirà nel punto B. Dato che la retta slitta da A a B, l'area dalla parte di A varierà in modo continuo da zero al massimo. Quindi c'è un punto tra zero e il massimo in cui l'area dalla parte di A è metà dell'area totale.

Abbiamo dimostrato per il piano un famoso teorema che si può generalizzare a tutti gli spazi di ordine superiore. Nello spazio a 3 dimensioni si possono dividere in due parti uguali con un piano i volumi di tre solidi qualsiasi; nello spazio a 4 dimensioni si può fare la stessa cosa con gli ipervolumi di quattro figure qualsiasi con un iperpiano a N-1 dimensioni. Il teorema, per lo spazio a 3 dimensioni, è detto «teorema del panino al prosciutto» perché si applica a un panino al prosciutto generalizzato consistente di due pezzi di pane e una fetta di prosciutto. C'è un piano che dimezza contemporaneamente tutti e tre indipendentemente dalla forma dei pezzi e da come sono disposti nello spazio.

La generalizzazione allo spazio N richiede tecniche matematiche molto po-



Forme da dividere in due parti uguali congruenti.

tenti, mentre per 2 e 3 dimensioni le dimostrazioni discendono facilmente dal nostro teorema fondamentale. Entrambe le dimostrazioni, in un certo senso, forniscono un modo per trovare una retta o un piano di bisezione, ma non è facile trovare praticamente tale bisezione. Non la si otterrà congiungendo i centri di gravità, perché una linea retta (o un piano) passante per il centro di gravità di una figura piana (o solida) non la dividono necessariamente in due metà di uguale area (o volume).

Data una figura semplicemente connessa, non necessariamente convessa, esiste sempre una linea retta che divide in due parti uguali contemporaneamente la area e il perimetro? La risposta è affermativa. La dimostrazione dell'esistenza di tale retta è molto simile all'ultima data. Si tracci una retta che divida in due parti uguali il perimetro in P e Q, come si vede nella figura in alto a pagina 126. Se tracciando questa retta si divide in due parti uguali anche l'area, non c'è altro da fare.

Ammettiamo che la retta non divida l'area. Si aggiunga alla retta una freccia rivolta dalla parte di area minore. Si spostino ora lungo il perimetro i punti P e Q nel senso delle lancette dell'orologio conservando sempre la bisezione del perimetro. Questa manovra costringerà la retta passante per P o Q a mutare in modo continuo il proprio orientamento, e contemporaneamente il rapporto delle aree da una parte e dall'altra della retta varierà in modo continuo. Dopo aver compiuto una rotazione di 180 gradi la retta riassumerà la sua posizione originaria, ma ora la freccia sarà rivolta verso la più ampia delle due regioni.

Si consideri la differenza tra l'area che sta dalla parte della freccia e l'altra area: all'inizio è positiva, alla fine è negativa. Questo valore chiaramente è una funzione continua dell'angolo della retta e quindi ci deve essere un angolo per cui il valore è zero e le due aree sono uguali. Il teorema può essere generalizzato anche a spazi superiori. Si può dimezzare il volume di un solido con un piano che ne dimezzi la superficie e, in generale, l'ipervolume di un solido di uno spazio N

Il Mago di Oz divide in due parti uguali un Mangaboo.

con un iperpiano N-1 che ne dimezzi la ipersuperficie.

Un altro bel teorema, ugualmente tutt'altro che ovvio, stabilisce che qualsiasi regione (anche questa volta non occorre che essa sia semplicemente connessa o neppure connessa) può essere esattamente divisa in quattro parti uguali da due rette perpendicolari. La dimostrazione standard è, a un tempo, sottile e meravigliosa.

Si immagini di avere un foglio di carta trasparente più largo della regione. Il foglio è diviso in quattro quadrati dalle rette perpendicolari X e Y. Il quadrante in alto a sinistra è grigio e quello in alto a destra è colorato. Si metta il foglio sopra la regione, con la retta orizzontale sotto la regione e quella verticale a destra. Si faccia scivolare il foglio finché X non divide in due parti uguali la regione e poi lo si faccia scivolare a sinistra (facendo scorrere X lungo se stessa) finché anche Y non divide la regione in due parti uguali. Si chiamino A, B, C e D (sulla regione, non sul foglio di carta trasparente) le quattro parti della figura. Si veda la disposizione delle lettere nella parte superiore della figura in basso della pagina seguente.

Per costruzione A più B è uguale a C più D e A più C è uguale a B più D. Sottraendo la seconda equazione dalla prima si ottiene $B—C = C—B$ ossia $2B = 2C$. Pertanto B è uguale a C e A è uguale a D.

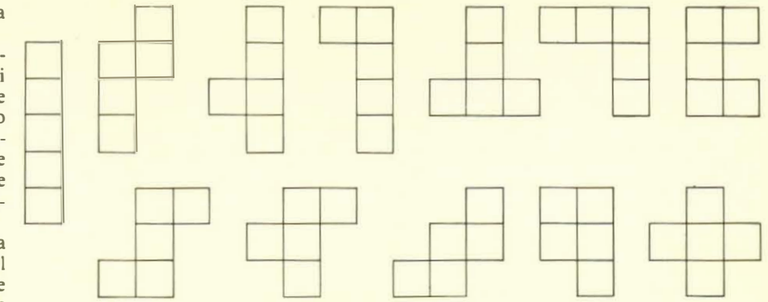
Se A è uguale a D, la regione è divisa esattamente in quattro parti uguali. Ammettiamo che non lo sia e che B sia più grande di A. Allora se si sottrae l'area della parte grigia della regione da quella colorata, si ottiene un numero positivo.

Si ruoti il foglio di carta di 90 gradi in senso antiorario, conservando sempre la doppia bisezione della regione operata da X e Y. Il loro punto di intersezione può vagare qua e là. Una volta completata la rotazione, le due rette ritornano necessariamente nella loro posizione iniziale, come si vede nella parte inferiore della figura, solo che ora X e Y si sono scambiate di posto.

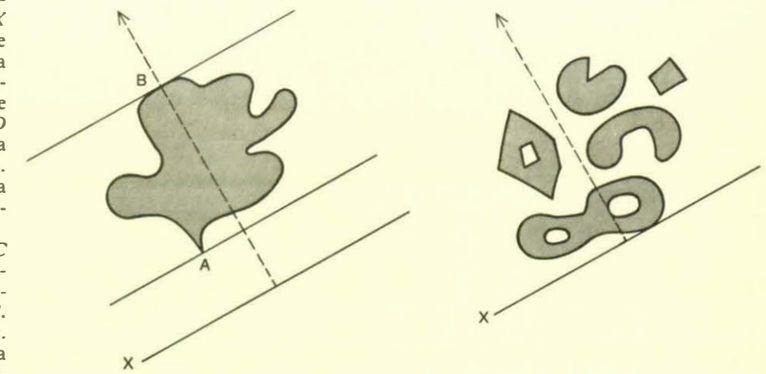
L'area colorata della regione è uguale ad A. L'area grigia è uguale a C e, dato che C è uguale a B, anche l'area grigia è uguale a B. Allora le aree della regione grigia e di quella colorata sono scambiate di posto. Di conseguenza, se sottraiamo l'area grigia, che è più grande, da quella colorata, che è più piccola, otteniamo ora un numero negativo.

È facile vedere come si applica il nostro teorema fondamentale. Il valore ottenuto sottraendo l'area della regione grigia da quella della regione colorata è una funzione continua della variazione dell'angolo di rotazione da 0 a 90 gradi. Dato che il primo valore va dal più al meno, ci deve essere un valore tra i due in cui la differenza è zero e le aree sono uguali. Quando si raggiunge questo valore le rette perpendicolari dividono la regione esattamente in quattro parti uguali.

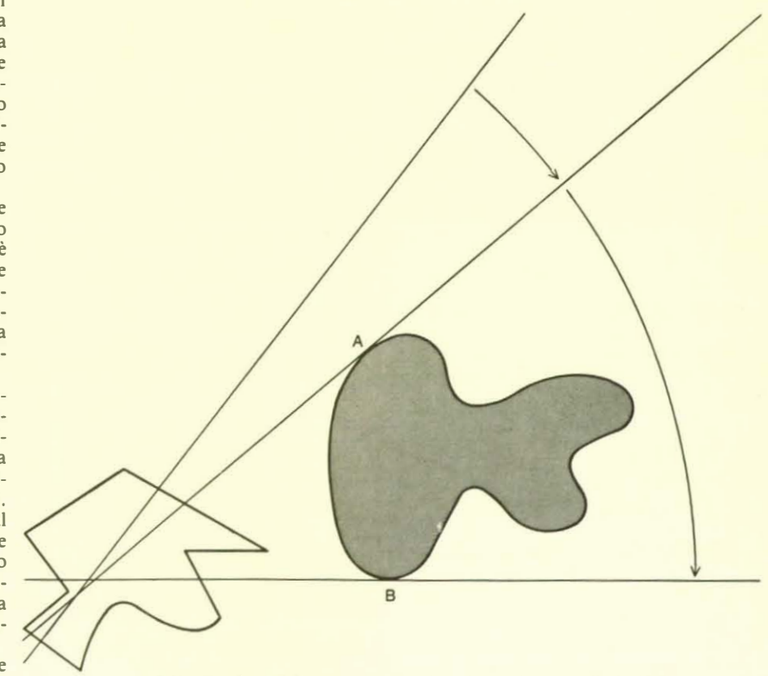
Questo teorema si può generalizzare anche a spazi superiori. Qualsiasi solido



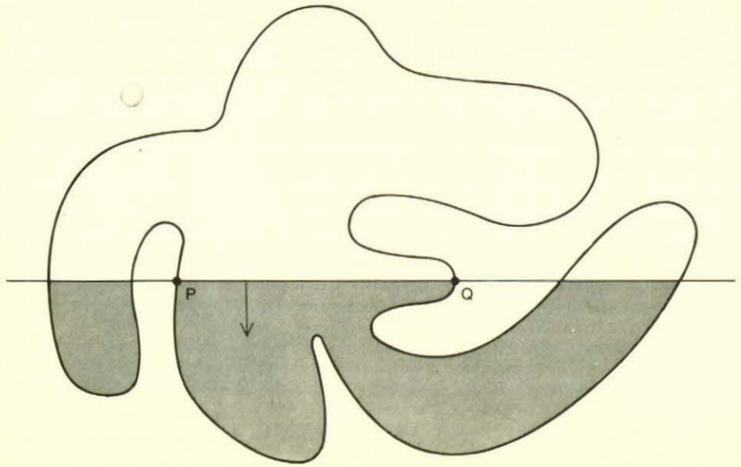
I dodici polimini di ordine 5.



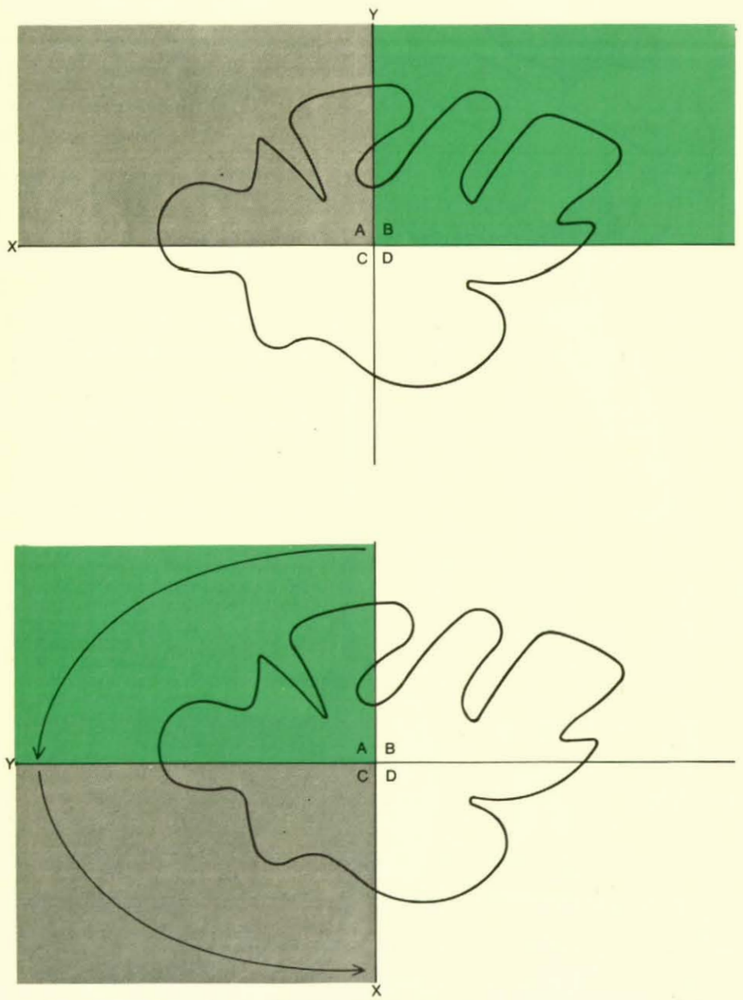
Dimostrazione di un teorema di bisezione dell'area.



Dimostrazione del teorema del panino al prosciutto per il piano.



Dimostrazione di un teorema di bisezione del perimetro e dell'area.



Come dividere una regione qualsiasi in quattro parti uguali con due rette perpendicolari.

può essere diviso in otto parti uguali da tre piani reciprocamente perpendicolari. Qualsiasi solido N-dimensionale può essere diviso in 2^n parti di uguale «volume» da n «piani» perpendicolari.

Anche qui, come per gli esempi precedenti il teorema non è di grande aiuto per il problema pratico della divisione in quarti di figure non simmetriche. Sappiamo, ad esempio, che un triangolo con i lati di 3, 4, 5 può essere diviso in 4 parti uguali da due rette perpendicolari, ma costruire effettivamente queste rette è tutt'altra cosa. In effetti non conosco nessun modo semplice per farlo.

La figura della pagina a fronte presenta quattro problemi di divisione in parti uguali, nei quali le costruzioni sono molto meno difficili di quanto non sia dividere in quarti il triangolo 3, 4, 5. Sono nondimeno abbastanza complicati da richiedere una notevole dose di abilità.

Il primo problema è dividere l'intera regione di nove quadrati in due parti di uguale area con una retta che passi per il punto P.

Il secondo problema è dividere in due parti uguali l'area totale dei cinque cerchi tracciando una retta per P, che è il punto al centro del cerchio più a sinistra.

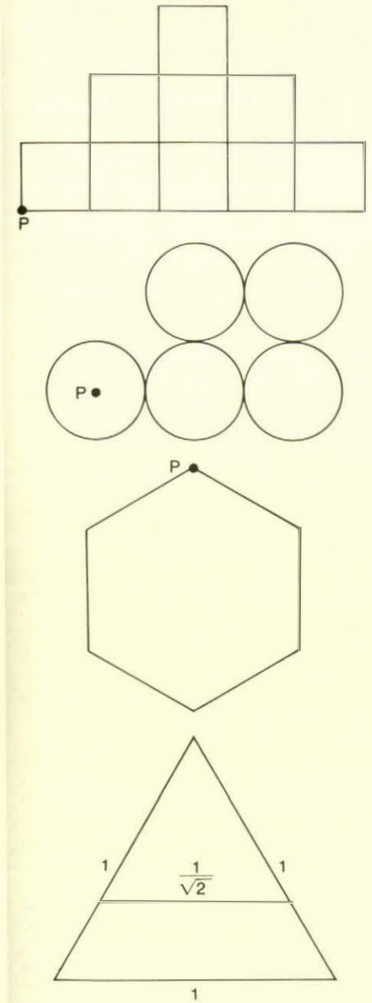
Il terzo problema è quello di dividere in tre parti uguali l'area di un esagono regolare con due rette passanti per P.

Il quarto problema è di dividere in due parti uguali un triangolo equilatero con una curva di lunghezza minima. La figura col triangolo mostra la più corta retta di bisezione, ma c'è una curva di bisezione più corta.

Di tutti e quattro questi problemi, come dei precedenti, verranno date le soluzioni il mese prossimo.

Il mese scorso, risolvendo uno dei problemi che riguardano i cappelli colorati, ho presentato una variante: ci sono n uomini, n-1 cappelli neri e uno colorato. Gli uomini sono progressivamente ciechi perché sono seduti in fila in modo che ogni uomo vede solo i cappelli davanti a sé. A, che è l'ultimo della fila, vede tutti gli altri cappelli, B li vede tutti tranne quello di A e così via fino all'uomo davanti che non vede cappelli. Gli n cappelli sono messi sulla testa degli uomini in un ordine qualsiasi e agli uomini viene chiesto (nell'ordine A, B, C...) se sanno il colore del proprio cappello. Lo uomo con in testa il cappello colorato sarà sempre in grado di rispondere affermativamente?

La soluzione è curiosa. Se A ha il cappello colorato, risponderà ovviamente di sì perché vede n-1 cappelli neri. Se B ha il cappello colorato dirà di no. Perché? Perché A deve sempre rispondere affermativamente, quindi il suo sì non dà nessuna informazione. Se C ha il cappello colorato risponderà di sì perché sente una risposta affermativa sia da A che da B. Se D ha il cappello colorato, deve dire di no perché la risposta affermativa di C potrebbe risultare dalla deduzione che lui, C, ha il cappello colorato. Questo processo si generalizza per induzione matematica fino alla conclusione c



Quattro problemi di divisione.

col cappello colorato dice di sì se e solo se è in posizione dispari a partire dal fondo. Se è in posizione pari, la risposta affermativa dell'uomo immediatamente dietro di lui non dà un'informazione sufficiente per una deduzione valida.

Nell'articolo del giugno scorso riguardante curve mostruose come le curve a fiocco di neve, le serpentine e le curve di Peano che saturano lo spazio si parlava di dimensioni «fratte», usando un termine di Benoît B. Mandelbrot, per curve con una dimensione di Hausdorff maggiore della dimensione del loro spazio topologico. *Fractals: Form, Chance and Dimension*, il libro che Mandelbrot ha scritto su queste curve e le loro molteplici applicazioni scientifiche, è stato recentemente pubblicato da W.H. Freeman and Company. È un bel libro abbondantemente illustrato con fantastici grafici ottenuti al calcolatore.